Detection of Anatomical Structures in Medical Datasets

Alison O'Neil, MEng

A themed portfolio submitted to
the University of Heriot-Watt

For the Degree of Doctor of Engineering in Optics and Photonics

Institute of Sensors, Signals and Systems

September 2016

Abstract

Detection and localisation of anatomical structures is extremely helpful for many image analysis algorithms. This thesis is concerned with the automatic identification of landmark points, anatomical regions and vessel centre lines in three-dimensional medical datasets. We examine how machine learning and atlas-based ideas may be combined to produce efficient, context-aware algorithms.

For the problem of anatomical landmark detection, we develop an analog to the idea of *autocontext*, termed *atlas location autocontext*, whereby spatial context is iteratively learnt by the machine learning algorithm as part of a feedback loop. We then extend our anatomical landmark detection algorithm from Computed Tomography to Magnetic Resonance images, using image features based on histograms of oriented gradients. A cross-modality landmark detector is demonstrated using *unsigned* gradient orientations.

The problem of brain parcellation is approached by independently training a random forest and a multi-atlas segmentation algorithm, then combining them by a simple Bayesian product operation. It is shown that, given classifiers providing complementary information, the hybrid classifier provides a superior result. The Bayesian product method of combination outperforms simple averaging where the classifiers are sufficiently independent.

Finally, we present a system for identifying and tracking major arteries in Magnetic Resonance Angiography datasets, using automatically detected vascular landmarks to seed the tracking. Knowledge of individual vessel characteristics is employed to guide the tracking algorithm by two means. Firstly, the data is pre-processed using a top-hat transform of size corresponding to the vessel diameter. Secondly, a vascular atlas is generated to inform the cost function employed in the minimum path algorithm. Fully automatic tracking of the major arteries of the body is satisfactorily demonstrated.

Acknowledgments

Thanks to my supervisors Dr. Ian Poole (TMVS) and Dr. Alex Belyaev (University of Heriot-Watt) for their unstinting guidance, support and shrewd criticism.

Thanks to the scientists at Toshiba with whom I collaborated on much of this work: Ian Poole, Sean Murphy, Brian Mohr, Mohammad Dabbah, Aneta Lisowska, Daniel Wyeth and Andy Murray. Thanks to Lynne McCormick and Prof. Graeme Houston of the Clinical Imaging University of Dundee, with whom I collaborated on work related to whole-body MRA vascular analysis and tracking.

Thanks to Corné Hoogendoorn and Keith Goatman who generously gave their time for proofreading and guidance on academic writing.

Major thanks to Erin Beveridge, clinical analyst at TMVS, who was ever helpful and knowledgeable in curating the ground truth and advising on the anatomical aspects of this work.

I would finally like to acknowledge Vital Images Inc. and the Clinical Imaging University of Dundee for providing some of the medical scans used in this thesis.

Contents

List of Tables

List of Figures

List of Acronyms

TMVS Toshiba Medical Visualization Systems

CT Computed Tomography

CTA Computed Tomography Angiography

MRI Magnetic Resonance Imaging

MRA Magnetic Resonance Angiography

ROC Receiver Operating Curve

LROC Localisation Receiver Operating Curve

AUC Area Under the Curve

2D Two-dimensional

3D Three-dimensional

HU Hounsfield Units

MIP Maximum Intensity Projection

MICCAI Medical Image Computing and Computer Assisted Intervention society

List of Publications

Publications on which I was an author during the course of this EngD are listed below.

Journal papers

Jim Piper, Yoshihiro Ikeda, Yasuko Fujisawa, Yoshiharu Ohnu, Takeshi Yoshikawa, Alison O'Neil and Ian Poole. Objective evaluation of the correction by non-rigid registration of abdominal organ motion in low-dose 4D dynamic contrast-enhanced CT. *Physics in Medicine and Biology* 57 (2012) 1701-1715.

Conference poster presentations

Alison O'Neil, Erin Beveridge, Graeme Houston, Lynne McCormick and Ian Poole. Arterial Tree Tracking from Anatomical Landmarks in Magnetic Resonance Angiography scans. *Proceedings of SPIE Medical Imaging 2014* 9034.

Alison O'Neil, Sean Murphy and Ian Poole. Anatomical landmark detection in CT data by learned atlas location autocontext. *Proceedings of the 19th Conference on Medical Image Understanding and Analysis* 189-194.

Patents

US 20140254906 Vascular tree from anatomical landmarks and a clinical ontology.
Inventors: Ian Poole, Colin Roberts, Paul Norman and Alison O'Neil.
Filed August 9th 2013. Granted 17th November 2015.

Chapter 1

Introduction

Abstract

In the introduction, we motivate the importance of detecting anatomical structures in medical scans, by the fact that this is an enabling technology for many common image analysis tasks. The phenomena that can make this a difficult task are outlined. We give a brief overview of the state of the art, setting the stage for more in-depth literature reviews in each technical chapter. Some commercial context is given to explain the investigative thrust of this thesis. We outline the main research questions that we will be attempting to answer, and give some overview of the structure of the chapters in this thesis and how they link together. Finally, a brief note on general scientific methodology is provided.

1.1 Why is detection of anatomical structures important?

In recent times, global healthcare spending has come under increasing pressure [2]. The cost of medicine is rising due to an ageing population [3], many technological advances — in drugs, surgical treatments and machinery —, international ambition to provide equitable healthcare regardless of wealth [2], and the emergence of a costly litigation culture [4]. Concurrently, the era of information technology has arrived. Imaging and non-imaging data is becoming plentiful and easily accessible, through mediums such as picture archiving and communication systems (PACS) and electronic health records. This brings practical and ethical challenges in ensuring data is fully examined and exploited on both an individual and a population-wide basis.

This is a world of many possibilities and high expectations but limited resources. The standout exception is computing hardware, which has become ever cheaper and ever more powerful, as famously predicted by Moore's Law. There is scope for medical image analysis algorithms to play a valuable role by aiding, automating and fail-proofing existing clinical tasks. Table 1.1 gives some examples of image analysis functions which can aid the routine work of radiographers, radiologists and surgeons.

All of these tasks could be facilitated, directly or indirectly, by detection of the anatomical structures in a scan. Points and regions of interest can be directly navigated to, labelled, segmented, enhanced and presented using relevant viewing modes. Segmentation allows measurement of distances, volumes and other quantities. Image registration using detected anatomical correspondences gives linking and fusion of different datasets. Identification of common anatomical sites for pathology assists detection of suspected lesions. Finally, metadata such as scan acquisition region and patient gender can be inferred from the anatomy that is present, or perhaps *not* present.

1.2 What makes identification of anatomy difficult?

Identification of anatomy is difficult due to intra- and inter-patient variation. *Intra*-patient variation occurs due to differences in patient pose, differences in phase of the respiratory and cardiac cycles, and differences in digestive system content. Long-term intra-patient variation occurs due to ageing, pathological

Image Analysis Task	Examples
Navigation of images	One-click navigation to a particular viewing plane or point location
Visualisation of images	Colour and illumination Curved multi-planar views (vessels) Open-rib view
Linking and/or fusion of images	Pre-contrast and post-contrast scans CT to MRI Imaging over time (follow-up)
Segmentation and labelling of particular tissues or organs	Heart chamber segmentation Vessel tree segmentation
Quantitative measurements	Arterial stenosis measurement Hippocampus volume
Locating regions of interest	Radiotherapy Scan planning
Detection of pathology	Lung nodule detection Mammography classification
Inference of patient and scan characteristics	Prediction of missing or erroneous DICOM tags
Non-invasive alternatives to current invasive techniques	Virtual colonoscopy
Decision making support	Imaging biomarkers as diagnostic tools or surrogate endpoints in clinical trials Big data analysis

Table 1.1: Examples of image analysis tasks

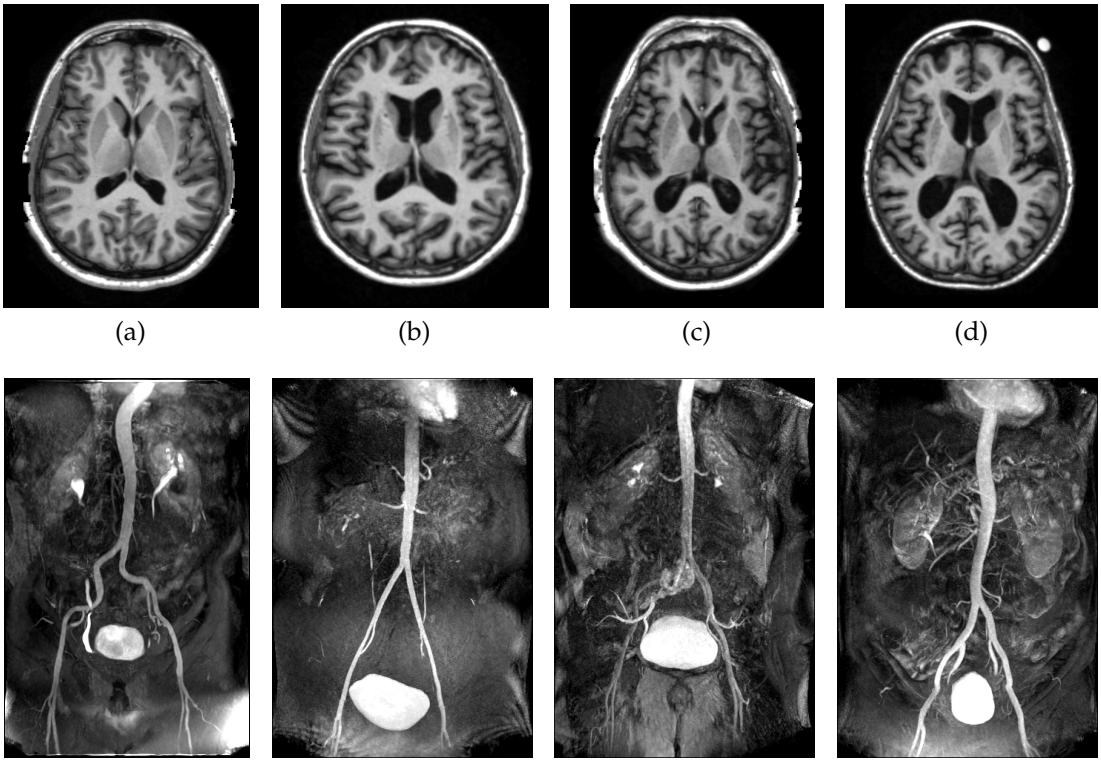


Figure 1.1: Illustration of inter-subject anatomical variation. Above: Equivalent axial slices from four MRI brain datasets. Below: MIP images from four MRA abdominal datasets. MRA Images © Clinical Imaging University of Dundee [TMVS Dataset IDs: 3571, 3590, 3591, 3592, 3740, 3754, 3759, 3771]

processes, and medical or cosmetic treatment. *Inter*-patient variation occurs because different people have vastly different sized and shaped anatomy, and even extra or missing structures. Figure 1.1 illustrates how anatomy varies from patient to patient. Scans may also contain imaging artefacts such as ringing, noise, bias field effects and motion blur. Later in the thesis, in Figure 2.5, there are plots of the types and frequency of variation which were observed in the 372 Computed Tomography (CT) datasets used for anatomical landmark detection.

The appearance of anatomical structures also changes dramatically depending on the imaging modality that is employed. There are many scanning technologies at the disposal of medical practitioners, the most common being X-ray, CT, Magnetic Resonance Imaging (MRI) and Ultrasound imaging. Even for the same technology, there will be differences between different scanners and makes of scanner, different protocols (for instance there are a vast number of possible MRI sequences), and differences in the usage and type of contrast agent. Generally,

scan resolution and extent is also restricted in order to limit the scanning time and, in the case of X-ray and CT, the radiation dose to the patient.

CT is possibly the most straightforward modality to deal with. Image grey level intensities are calibrated according to the Hounsfield Units (HU) scale [5]. There are specific values for different tissues which are approximately: Water 0 HU, Air -1000 HU, Bone 1000+ HU, Fat -50 HU and Muscle 40 HU.

Other modalities have uncalibrated grey level intensities. Even where images have similar appearance to the human eye, the data range and distribution is often quite different, both scan to scan and even within a scan e.g. bias field effects or contrast perfusion variation.

1.3 What is the state of the art?

Generic image processing techniques such as thresholding, filtering, histogram analysis, morphology and region growing have long been used for segmentation or enhancement of different tissues.

However, intelligent identification of particular named anatomical structures really became possible with atlas registration — be it with a single atlas, multiple atlases or a probabilistic atlas. The term *atlas* refers to a reference image which a human observer has labelled with the anatomical regions of interest. The variation between scans makes perfect one-to-one correspondence difficult. Nonetheless, image registration works well in body parts such as the head where the range of natural variation is somewhat constrained. Correspondences are found either by direct matching of grey level intensities or matching of image features derived from the intensities. Simple affine registration seeks a basic *global* mapping between images. In deformable registration, images are warped on a *local* basis to give a flexible mapping with many degrees of freedom. Deformable registration is a powerful technique; the main drawback is that state-of-the art algorithms such as ANTS-Syn [6] have run times of the order of many minutes or even hours [7, 1]. See [7] for an evaluation of popular deformable registration algorithms.

Atlas techniques are also a possible approach for landmark detection [8, 9]. Alternatively, there are template matching methods [10] where the volume is searched to find the best matches for a set of template patches. Similarity metrics are used to identify matches, as in whole-volume registration. Statistical shape modelling techniques [11, 12, 13] where an appearance model is combined with a statistical shape model is another well-known approach, with a common application being the matching of surface landmarks in segmentation tasks.

As increasing amounts of data and computing power become available, machine learning techniques have come to the fore in the world of computer vision. These techniques aim to discover a useful model directly from the data, rather than said model being explicitly provided. Given a large amount of training data, machine learning algorithms can be accurate, robust and fast. Accuracy comes from the ability to model complex, non-parametric data distributions. Robustness comes from the ability to learn salient, generalisable data characteristics. Speed comes from the fact that the majority of the work is done off-line during the learning phase. The parameters of a machine learning algorithm are learnt automatically during the training stage; the skill comes in choosing the *hyperparameters* of the algorithm and designing data augmentation schemes when ground-truth is scarce. Examples of machine learning algorithms are support vector machines, random forests and neural networks. In particular, random forests have been demonstrated as a solution for many anatomical detection problems [14, 15, 16, 17, 18, 19, 20].

In this thesis, we join other researchers [21, 22, 23] in exploring the intersection between the fields of image registration and machine learning. The goal is to leverage the advantages of each of these approaches, which are driven by contextual and feature information respectively.

1.4 Medical image analysis in a commercial environment

Toshiba Medical Visualization Systems (TMVS) is a subsidiary of Toshiba Medical Corporation Systems (TMSC). TMSC is one of the big four in the market for medical scanning and visualization systems, along with Siemens, GE and Philips [24]. As a commercial company, it is important that technology is competitive and clinically viable.

- Run times must be practical for doctors to tolerate as part of a normal clinical work flow.
- Accuracy should be sufficient for diagnosis and/or for surgical precision.
- Systems must be robust. There is limited use for algorithms which work only on healthy subjects, or are prone to failure.
- There are hard limits on the availability of memory and computing power.

Tuning of systems for optimum performance is a major component of the image analysis work that is done at TMVS, to find the best trade-off between the factors outlined above. Any algorithmic complication must be justified in terms of the performance benefit that it confers. Thus, in many places in this thesis a comparison is presented between the simpler and the more complicated method, or an analysis is presented of the trade-off between accuracy and run time performance.

In order to ensure robustness, all of the code is comprehensively unit-tested, and good software engineering practice is important. In accordance with this philosophy, we try to extend *existing* mature and well-tested algorithms where possible. In particular, we take a random classification forest algorithm for landmark detection in CT [25], a multi-atlas segmentation algorithm for the brain [26] and a two-point vessel tracking algorithm based on finding the minimum cost path, similar to that of Kanitsar *et al.* [27].

1.5 Research questions

This thesis addresses the problem of finding commercially viable solutions to the following questions:

- How may spatial relationships between landmarks be exploited in a machine learning algorithm for anatomical landmark detection?
- How may an existing random classification forest for anatomical landmark detection in CT data be adapted for use with other imaging modalities?
- How may existing multi-atlas segmentation and random forest classifiers be combined for the problem of brain region segmentation?
- How may we develop a fully automated system for tracking and labelling the major arteries of the body in Magnetic Resonance Angiography (MRA) datasets?

1.6 Thesis overview

A diagrammatic overview of the content of this thesis is given in Figure 1.2.

We start in chapter 2 with the problem of anatomical landmark detection. Figuring out how to reliably detect a variety of landmarks on different tissues

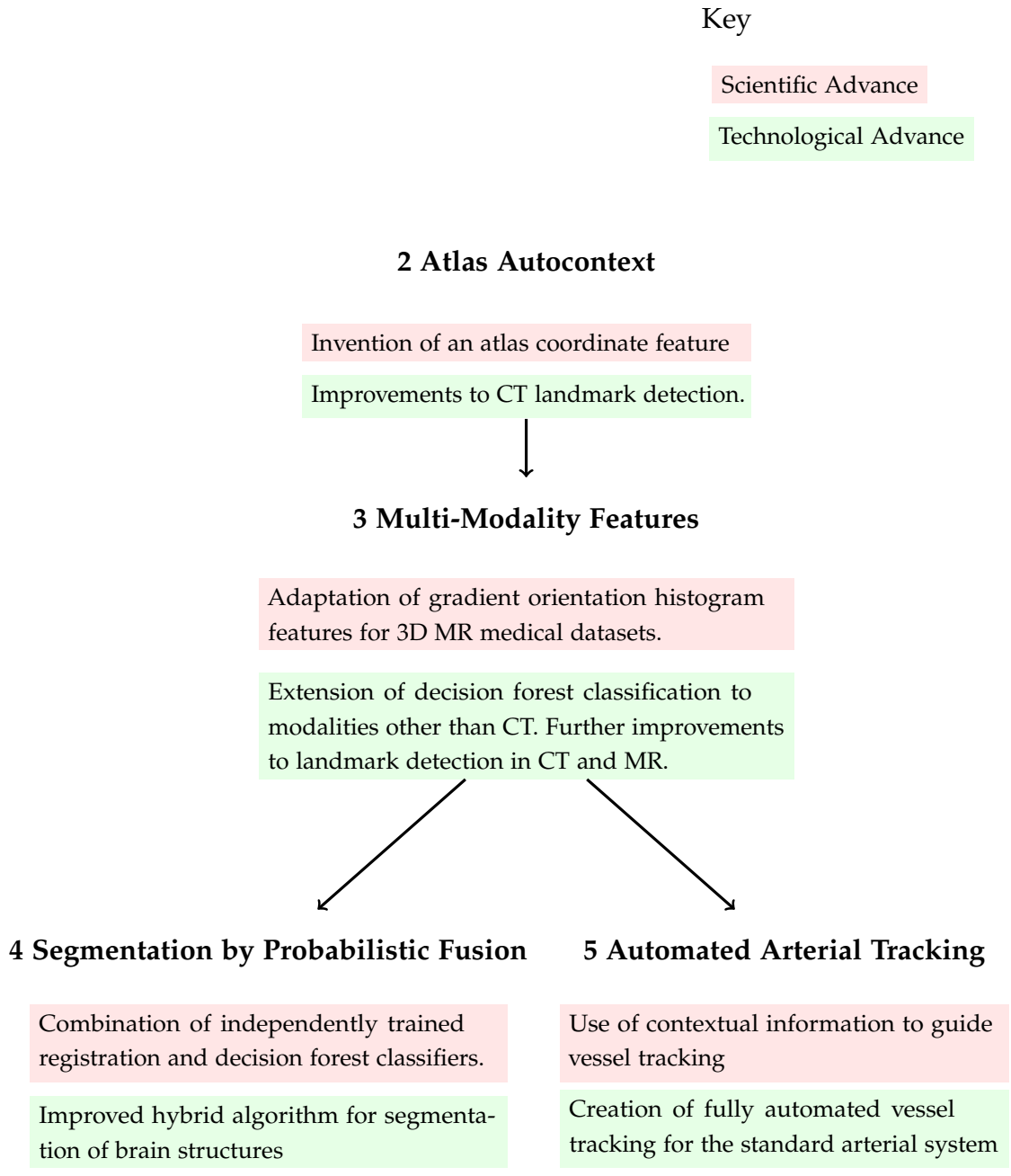


Figure 1.2: Overview of thesis content by chapter, split into the scientific advances that we worked on (red) and the practical technological advances that these enabled (green).

and body systems is relevant since many algorithms can be initialised or augmented with the use of landmarks. We examine how to incorporate learnt spatial information, developing an analog to the idea of *autocontext* [28], termed *atlas location autocontext*, whereby spatial context is iteratively learnt by the machine learning algorithm as part of a feedback loop.

Chapter 3 looks at how to extend our landmark detection capability from CT to MRI. We are interested in how to design image features for uncalibrated medical data, or even for unseen modalities. With this in mind, an investigation of gradient orientation features — as used in the HOG and SIFT descriptors — is presented. We examine the dimensions of the cuboids over which the histogram is computed, the random sampling strategy for the cuboid offsets, the number of histogram bins, the plane in which the Two-dimensional (2D) gradient orientations are measured, noise thresholding and weighting schemes. We then evaluate the effectiveness of gradient orientation features compared to the simple intensity features that have been used previously by scientists at TMVS. Cross-modality landmark detection is demonstrated using *unsigned* gradient orientations for MRI-T1, MRI-T2 and CT data.

In chapter 4 the random forest classifier is adapted for the purpose of brain gyrus segmentation. We compare this classifier to a multi-atlas segmentation method, and examine simple methods of combining the two classifiers. The results are benchmarked against other algorithms using data and participant results from the MICCAI 2012 Grand Challenge on multi-atlas labelling of the brain [1]. We show that the hybrid algorithm gives good results in terms of speed and accuracy.

Finally, in chapter 5 we present one application of anatomical landmarks: vessel centre line tracking. A set of vascular landmarks is defined which allows us to track the arterial tree, and we optimise the landmark classifier for vascular landmarks. An existing vessel tracking algorithm is then used to track between landmarks. We develop mechanisms for augmenting the tracking algorithm with contextual information in the form of vessel-specific filters and atlases.

1.7 Scientific methodology

In chapters 2 and 3, where we look at optimisation of the random forest classification algorithm, the forest is trained three times for each experiment using three different random seeds and all three results are plotted. This is to give an idea of the degree of performance variation inherent in the method, such that consistent

improvement (or indeed regression) can be distinguished from random noise.

Chapter 2

Anatomical landmark detection by learnt atlas location autocontext

Abstract

This chapter introduces the problem of anatomical landmark detection. The problem domain is CT data, and scans are considered from any scanner and scanning protocol, acquired for any part of the body. Image intensity values in CT scans are calibrated and therefore local neighbourhood intensities are descriptive and discriminatory features for a given location. A pre-existing random forest approach is described which learns to detect landmarks on the basis of intensity features. This algorithm has the flaw that landmarks may be detected in spatially inconsistent places; sometimes landmarks are detected even though they are outside the region of acquisition. In an extreme case, an abdominal landmark may be detected in a head scan. Hence, we build on this algorithm by introducing a mechanism, atlas location autocontext*, to exploit knowledge of spatial context. Atlas location autocontext is a feedback mechanism, where the results of an initial classifier are mapped to a landmark atlas, and the resulting estimated x, y and z atlas space coordinates for each sample location are used as features in addition to the local intensity features to train a second spatially-aware classifier. This process iterates, such that classifiers with increasing spatial awareness are trained, and in this way the results are refined to correct those landmarks which are initially detected in spatially inconsistent positions. In practice, we find that two feedback loops give maximum benefit. A simple affine transformation gave best performance for the atlas mapping that we fed back, confounding our hypothesis that a non-rigid spline mapping would be the more biologically accurate and therefore more effective mapping. We explain this by the fact that in regions where landmarks are detected poorly, the affine transformation in fact has error cancellation properties where errors are distributed evenly. Even if not, the global transformation will be subject to only mild distortion as a result of a wrongly detected landmark, in contrast to the spline which is flexible enough to map wrongly detected positions exactly. We finish by describing the residual sources of error, some of which (poor landmark definitions and too low operating resolution) will be addressed in chapters 3 and 5.*

2.1 Synopsis

In this chapter we

- (2.3.2) Describe a pre-existing random classification forest approach to landmark detection.
- (2.3.3) Introduce *atlas location autocontext*, a lightweight spatial context mechanism suitable for use with machine learning algorithms as applied to landmark detection problems. This consists of feeding the estimated atlas coordinate of each voxel as a feature to a subsequent detector, and iterating to give a sequence of detectors yielding progressively more accurate results.
- (2.3.3.3) Evaluate the accuracy of different rigid and non-rigid atlas mapping functions, showing that the thin plate spline is most accurate.
- (2.4) Show that atlas location autocontext gives improvement for the first two iterations feedback, and that thereafter the results are stable.
- (2.4) Show that an affine transform gives greater improvement in accuracy of landmark detection, in terms of mean error and Area Under the Curve (AUC), than a thin plate spline when employed for atlas location autocontext.
- (2.4.3) Analyse the effect of atlas location autocontext on individual landmarks.
- (2.5.2) Explain the good performance of the affine mapping in terms of its error-cancellation property.
- (2.5.3) Investigate the reasons for the remaining landmark errors, namely: similar looking repeating structures (ribs and vertebrae), anomalous datasets for which we have few examples, poorly defined landmarks, and landmarks on fine structures which are visible only at a higher resolution than that at which we run the detection algorithm (in particular, small-vessel landmarks).
- (2.6) For future investigation, suggest improvements to the atlas mapping, define some derivative features which leverage the atlas location concept, and introduce the idea of using a prior via the *atlas location autocontext* mechanism.

2.2 Introduction

2.2.1 Problem description

This chapter is concerned with the detection and localisation of anatomical landmarks in medical CT scans. We consider *named* anatomical landmarks, as opposed to identifying anonymous distinctive points. Hence, a landmark is a point location which can be described in terms of the anatomical landscape. A few examples are "superior aspect of right eye globe", "posterior tip of spinous process of C7", "head of pancreas" and "bifurcation of trachea". See appendix A for a complete landmark list. The distribution of the skeletal landmarks is shown in Figure 2.1.

Figure 2.1: Schematic of the skeletal landmarks (soft tissue landmarks are omitted for clarity). Landmarks are indicated with purple dots. It can be seen that the landmark set has coverage over the whole body, excluding the upper limbs.

We define a set of landmarks covering all body systems and regions, however the method must work for CT scans of any part of the body. Therefore, our algorithm must both *detect* (is it there?) and *localise* (where is it?) each landmark.

Landmark detection is an important enabling technology, since landmarks can be used to initialise or augment many algorithms, for instance deformable volume registration [29, 30, 31, 32, 33, 34, 35], organ segmentation [36, 37, 38, 39, 18] and vessel tracking (see chapter 5).

2.2.2 Prior art

Early techniques to identify anatomical landmarks were based on a prior belief as to their appearance. In 1997, Rohr extended a number of 2D operators to three dimensions for the detection of point landmarks [40]. Soon after, Wörz and Rohr [41, 42] developed geometric models for representing tip-like, saddle-like and sphere-like structures based on the ellipsoid equation. Extra parameters were

incorporated to model image blurring (a Gaussian smoothing operation with width parameter σ) and spatial transformations (translation, rotation, tapering, bending). These schemes were demonstrated on medical Three-dimensional (3D) CT and MRI volumes. A less abstract approach was taken by Betke *et al.* [10] who presented template-based discovery of landmarks in chest CT scans. Template images — one per landmark — were manually cropped from a training dataset and then matched to the test dataset by finding the patch best correlated with each template.

Spatially-driven landmark matching can be attempted using global registration to an atlas. In the same way that there is a close relationship between segmentation and registration [43], so is there a close relationship between landmark detection and registration. Perfect registration would implicitly yield perfect landmark detection, and vice versa. Atlas-based approaches have exploited this. For instance, Ehrhardt *et al.* [8] used Demons-based non-rigid atlas registration [44] to identify bony hip landmarks. Iglesias and Karssemeijer [45] used a multi-atlas algorithm to detect landmarks in mammograms. Liu and Zhou [9] performed a search over 10,000 volumes to find the closest match to the novel volume, and then directly transferred the landmark annotations across, followed by refinement using local landmark detectors. In active shape models, an appearance model is combined with a statistical shape model, and the space of possible landmark configurations is searched to find the most likely [11, 12, 13, 46].

However, state-of-the-art atlas registration is a slow process, and is vulnerable to inter-subject variability. More recently, machine learning techniques have gained in popularity for segmentation and landmark detection problems, for a number of reasons. Firstly, arbitrary locations of any appearance may be selected as landmarks (as opposed to constraining to e.g. edge locations or tip-like locations) since salient features are discovered automatically from the training data. Secondly, given enough training examples, it is possible to achieve high accuracy. Finally, the majority of the work is done off-line during the learning phase, resulting in fast detection at the application stage.

A popular machine learning option is random decision forests, pioneered by Breiman [47]. They were used by four out of five entrants in a 2015 MICCAI challenge for landmark detection in cephalometric X-ray images [48]. Random forests are an expansion of classification and regression trees into ensembles of many decision trees, for which the results are aggregated. In random forests the idea of *bootstrap aggregating* [49], or bagging, is employed so that each tree sees only a portion of the training data. Further, the *random subspace method* [50] may

be employed to apportion a randomly selected subset of the images feature vector for each tree. A further technique is the random split method [51], where at each node of the tree, a split is selected at random from the K best splits according to the information gain. All of these techniques aim at randomising the training of the trees such that they are randomly different to one another. This decorrelation decreases the variance of the forest classifier, compared to the variance of a single decision tree [49], giving good generalisation to unseen data. In a classification forest, each tree yields an estimated class probability distribution for a given novel sample. In a regression forest, each tree yields estimated distributions for the continuous output variables (i.e. in landmark detection, the variables are typically the displacements of the test voxel from each landmark).

By default, landmarks are detected independently. This can lead to some anatomically implausible landmark configurations. For instance, it is irregular to detect an abdominal landmark superior to a thoracic landmark, or to detect the L2 vertebra superior to the L1 vertebra. The subject of exploiting contextual information to improve landmark detection accuracy has been tackled by a number of authors.

In the *entangled forests* of Montillo *et al.* [15], as each tree is grown, information is gathered about the path of traversal and estimated class probability distribution for each voxel and its neighbours, and this provides feature values to deeper nodes in the tree. At classification time, voxels are classified simultaneously and each voxel (may) influence the path of its neighbours. This technique was demonstrated in the segmentation of anatomical organs in CT images. In *structured random forests*, Kontschieder *et al.* [52] measure information gain jointly over the sample voxel and its neighbours, so that similar label patches are clustered within the leaves. Leaves store probability distributions for the whole patch, thus generating overlapping patch label predictions during classification.

An alternative approach is to infer context from the results of an initial classifier. This is the idea behind *autocontext*. A second classifier is trained on the voxelwise class probabilities returned by the first classifier, with or without knowledge of the image data. The process is iterated so that a sequence of classifiers is trained, leading to progressive refinement of the results. This technique was first demonstrated in satellite images by Poole [53], and later for segmentation of structures in brain MRI by Morra *et al.* [28] and by Tu and Bai [54]. In landmark detection, class-related contextual information is sparsely distributed due to the point location nature of the landmarks, which are surrounded by background class samples. This makes autocontext less effective.

Graph matching techniques have been demonstrated as a post-processing step to select the combination of landmark candidate locations according to *geometric* constraints. Landmarks are designated as nodes, and edges designate connections between spatially correlated landmarks. Donner *et al.* [16] formulated such a Markov Random Field (MRF) graphical model which modelled relative pairwise landmark displacements as Gaussian distributions. Landmark candidates were first identified using random forests. The best combination of candidates was then identified by solving the MRF, taking into account the forest classification probabilities (as unary potentials), and displacements between pairs of candidates (as pairwise potentials). Guo *et al.* [17] went further, modelling both landmark-landmark distances and *intensity profiles* along the straight-line landmark connections, by measuring the cross-correlation in the so-called *line patches*, for the detection of landmarks in hand X-ray images.

Aided by the use of a *regression* forest for locating landmarks, Gao and Shen [18, 55] took an approach which combines the ideas of geometry and feedback of results, for the problem of landmark detection in prostate and head & neck CT images. Landmarks were explicitly separated into reliable and dependent groups. A two-layer implementation was then trained in which the first layer yielded estimated displacements from each voxel to each of the *reliable* landmarks, and the second layer used these as features in addition to standard image features to detect all landmarks. Each layer was multi-resolution, containing regression forests from coarse to fine resolution (the search space also being progressively decreased).

2.2.3 Motivation behind our approach

Our work builds on that of Dabbah *et al.* [25] of TMVS, in which the random classification forest of Criminisi *et al.* [14] for finding abdominal organ bounding boxes was adapted for the problem of finding landmarks in miscellaneous CT scans. We have a mature implementation for the algorithm (see section 2.3.2 for details).

We propose a multi-layer decision forest similar to that of Gao and Shen [18, 55] but we unify the spatial information yielded from all detected landmarks into a set of three *atlas location* features comprising the x, y and z estimated atlas space coordinates. This is a higher-level spatial feature which can be derived from the results of any initial detection algorithm. After distilling the probabilistic output of the forest into a set of confidently detected landmark positions, we

register the set of landmarks to a whole-body landmark atlas and feed back the estimated atlas space mapping to a subsequent random forest detector. We term this technique *atlas location autocontext* because of its similarity in principle to *autocontext*; the difference is that in our approach we feed back a learnt spatial mapping rather than a learnt set of voxelwise class probability distributions. Atlas space is robust to inter-patient differences in scale, orientation and the acquisition area of the scan, since these differences can be easily accounted for in the mapping to atlas space by simple scaling, rotation and translation operations. Further, contextual information is unified across the volume rather than considered to be independent for each voxel. Using this representation, we may easily discard landmark results in which we have low confidence by excluding them from the mapping.

Atlas location autocontext uses significantly less memory than autocontext, where the full set of class probabilities must be retained and fed back for every voxel in the volume. Rather, contextual information is condensed into a sparsely parameterised mapping function, making this an efficient feedback mechanism.

A random classification forest is used, however our contribution is relevant to other machine learning methods.

2.3 Method

An overview of the landmark detection method is given in Figures 2.2 and 2.3, which show the training and detection phases respectively.

We start by outlining some characteristics of the data cohort that we are working with. We describe the (pre-existing) random classification forest that is used as the base detector. We then describe the proposed atlas location autocontext mechanism, and finally present the results of an investigation into the accuracy of different mappings to atlas space, from which we select the best for use in our landmark detection experiments.

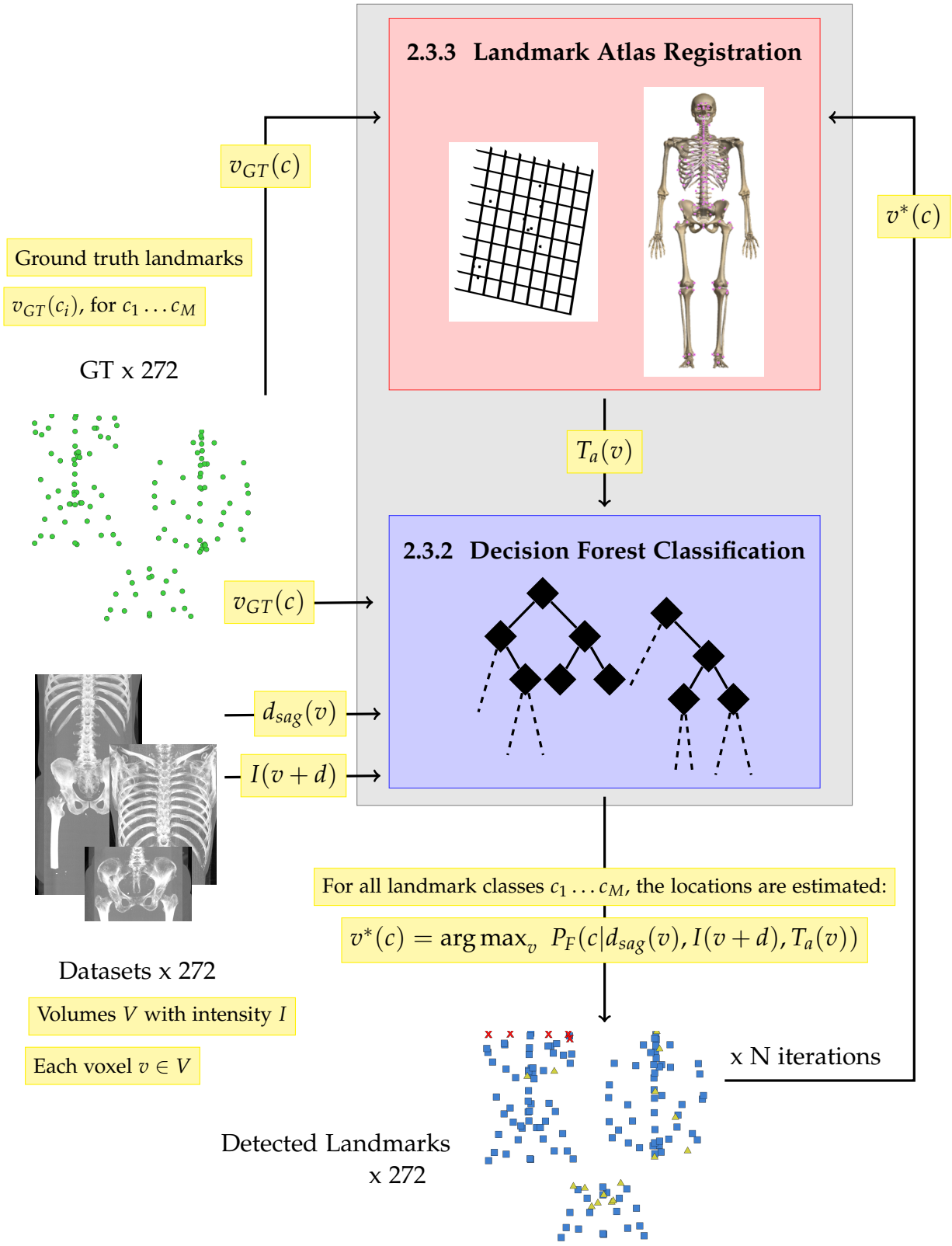


Figure 2.2: Overview diagram of the training phase of the anatomical landmark detection algorithm. In this chapter we focus on the computation and use of the learnt atlas location $T_a(v)$ as an autocontext machine learning feature (step marked in pink), in addition to the intensity $I(v+d)$ and sagittal displacement $d_{sag}(v)$ features. Section 2.3 has full details.

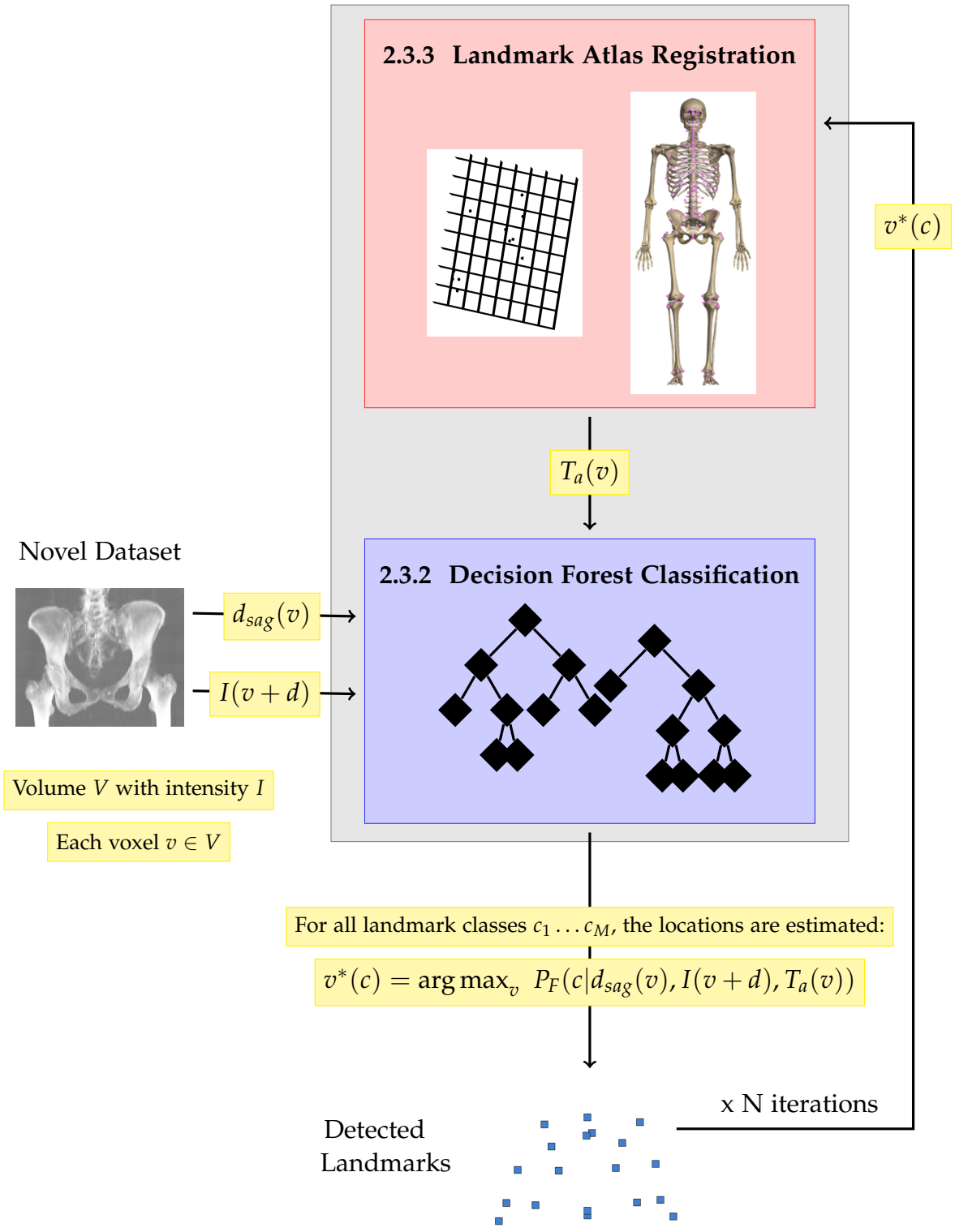


Figure 2.3: Overview diagram of the detection phase of the anatomical landmark detection algorithm. Section 2.3 has full details.

2.3.1 Data and ground truth collection

2.3.1.1 Datasets

A cohort of 452 CT datasets were used. These came from multiple scanner vendors, inclusive of both contrast and non-contrast acquisitions, and covering a range of anatomical regions and acquisition volumes including brain, head & neck, thorax, cardiac, abdomen, full-body and lower limbs.

We tried to make the mix of datasets representative. To this end, half of the head cases were removed since the head region was over represented. A few highly anomalous datasets were excluded as in the original work [25]. Aside from these omissions, all datasets to which we have access have been used without prejudice. The remaining 372 datasets were split into 100 test datasets and 272 training datasets. The test-train split was done by random selection, stratified with respect to body region.

Figures 2.4 and 2.5 illustrate the distribution of different attributes of the data population. There are histograms of gender, body part, scan manufacturer, presence of contrast, and presence of variation, be it medical, anatomical or pathological. The test-train split was done by random selection, stratified with respect to body region, and this is evident from the body region graph, which has a similar shape for the two populations. The large number of instances of medical and pathological anomalies is unsurprising given the clinical indications for performing a CT scan.

Figure 2.6 shows a scatter graph showing the different dataset voxel resolutions. These plots indicate that the test and training cohorts are similar. Most datasets have lower resolution in the Z-direction. In a few cases this is lower than the 4mm voxel^{-1} scale that we employ during landmark detection; this rather coarse scale was found to give best accuracy in cross-validation experiments (see section 2.3.2.7 for a description of parameter values).

We can see that altogether we have a diverse dataset population. The intention is to design an algorithm which is robust to the scanner protocol and patient characteristics, such that it could be routinely run on all performed scans.

2.3.1.2 Ground truth collection

Ground truth data was manually collected by anatomically trained colleagues at TMVS. For the work in this chapter, a set of 127 landmark positions was defined as given in appendix A. Landmarks were chosen on the basis of clarity of

location, clinical utility, and in the interests of giving good anatomical coverage. Gender-specific and anatomically atypical landmarks were avoided. A schematic of the skeletal landmarks is shown in Figure 2.1. Soft tissue landmarks are omitted from this diagram for clarity.

During collection of ground truth, the placement of ground truth was sometimes a matter of debate:

- **Soft tissue landmarks:** Soft tissue landmarks are not always visible in CT scans e.g. the pancreas extremities may be hard to see.
- **Vascular landmarks:** Vascular landmarks are difficult to see in non-contrast scans.
- **Ribs and vertebrae landmarks:** Where only a central portion of the thorax is visible, for cases of repeating structures such as ribs or vertebrae, the marker cannot count from either the top or the bottom instance and hence can only be sure to an accuracy of one or two which is which (by looking at the relative positions of soft tissue organs and other bony structures).
- **Landmarks obscured by surgical or pathological variation:** The landmark may be obscured by e.g. tumours, vessel stents, joint replacements.

In these cases, the marker made their best guess at the landmark positions but marked them as uncertain. We then excluded uncertain landmarks from the numerical results, on the basis that since a human observer cannot be sure of the landmark position (or even whether it is present in the scan) we do not place any expectations on the automated algorithm. Uncertain landmarks are in evidence in the pictorial results that we show later (in particular, see the vertebra in 2.20 and the pancreas head in 2.21).

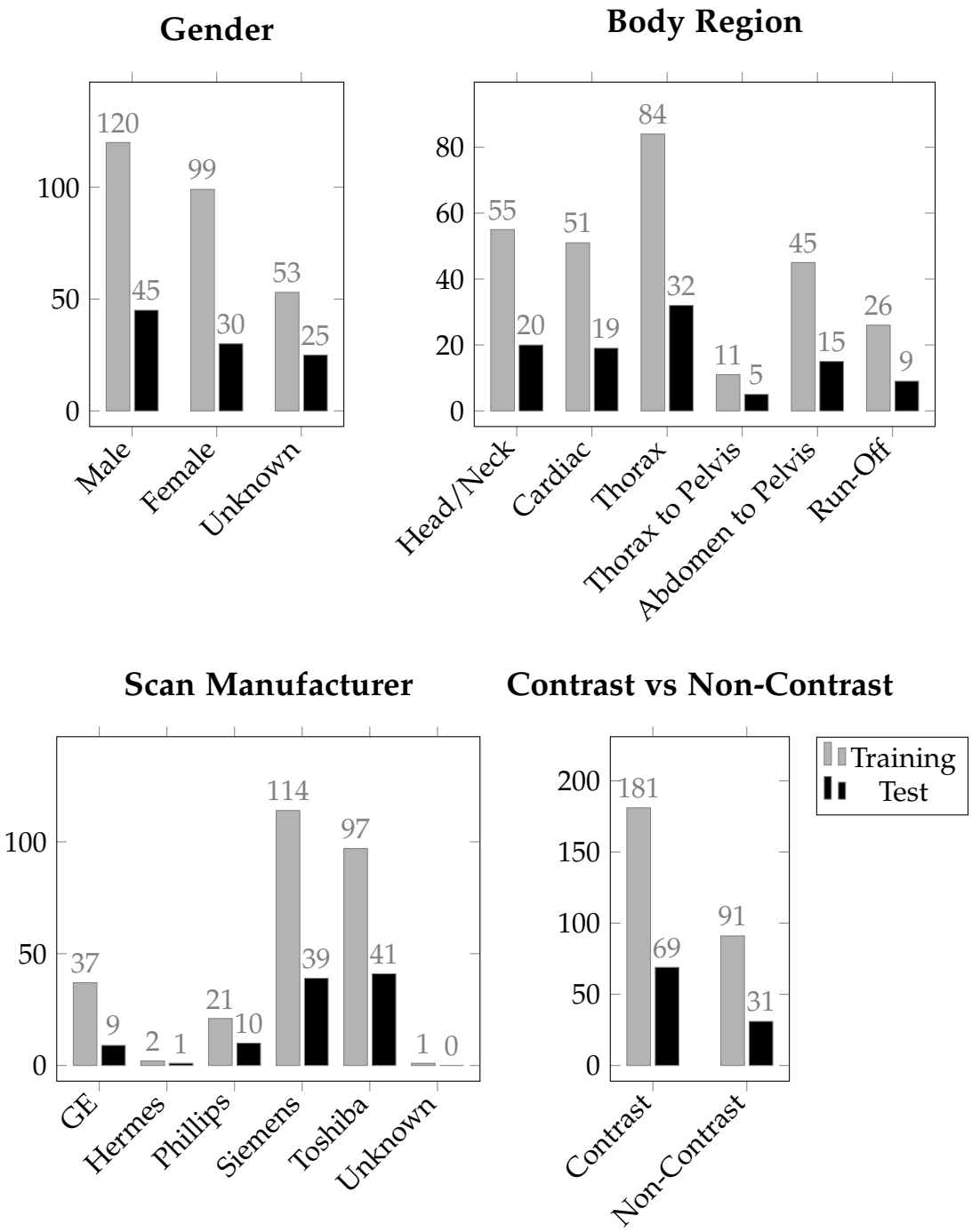


Figure 2.4: Plots showing data distributions of gender, body region, scan manufacturer and presence of contrast. The vertical axis shows the frequency of datasets. Gender was unknown for scans in which the gender had been anonymised and gender-specific organs were not visible. Run-off scans generally show the lower limbs and some or all of the upper body.

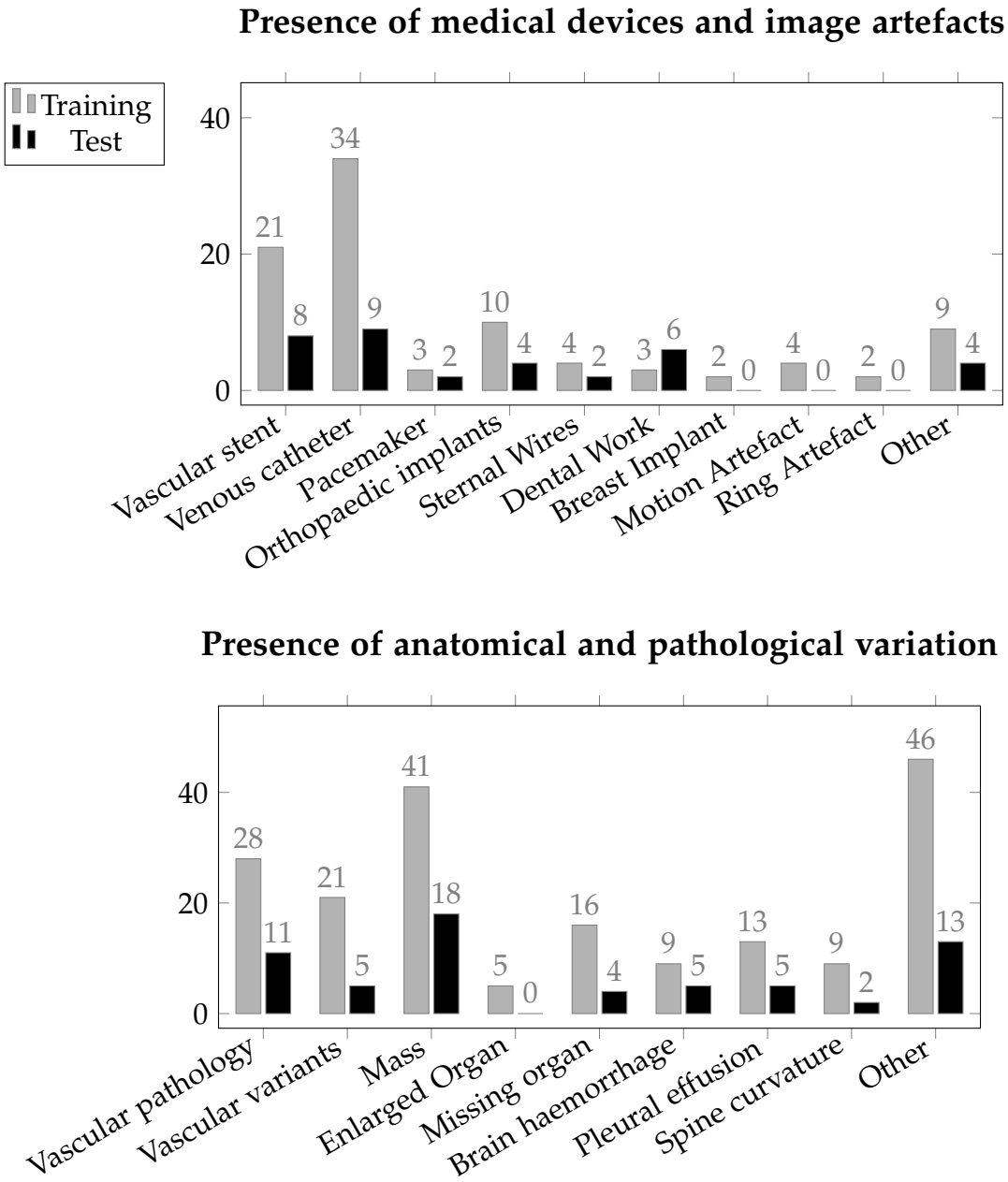
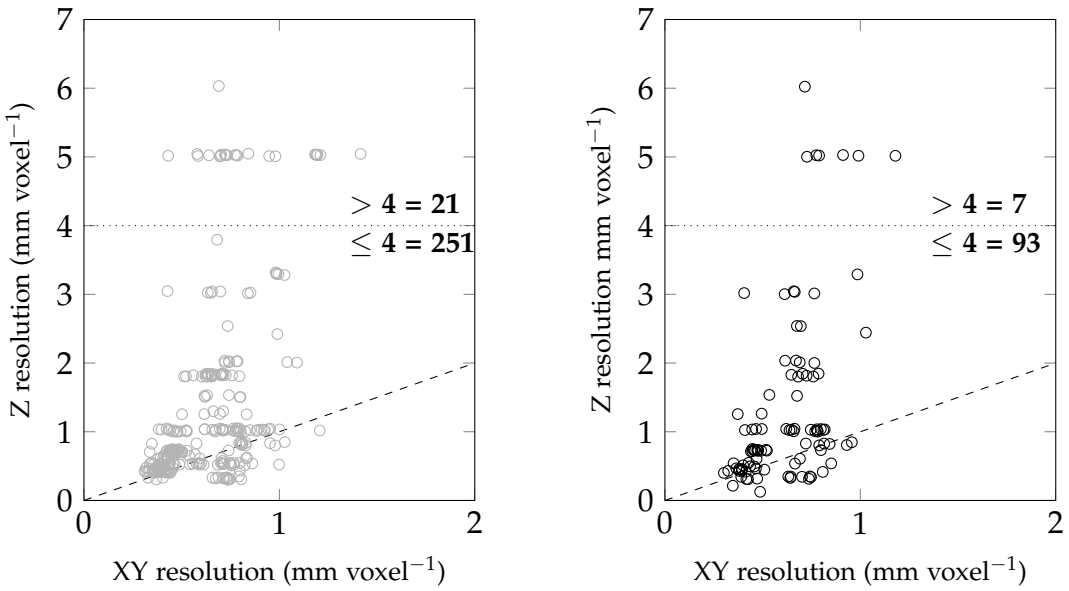


Figure 2.5: Plots showing frequencies of data variation due to medical instrumentation, imaging artefacts, anatomical differences and pathology. The vertical axis shows the frequency of datasets. "Orthopaedic implants" refers to joint replacements and metal plates/pins. "Mass" refers to tumours, nodules and cysts. Variation was recorded by the anatomists during manual collection of ground truth, who noted cases where pathological or other variation was either very obvious or interfered with the placement of landmark points. Hence these figures are not an exhaustive or clinically certified inventory. Each dataset may contain more than one instance of variation, but only one instance of each category is counted per dataset e.g. a bilateral hip replacement counts as 1 instance.



(a) Training

(b) Test

Figure 2.6: These (jittered) plots show the distribution of resolution values in the training and test data cohorts. Jitter has been added to the points because there are cases where many datasets have the same resolution. The mm voxel^{-1} in the X and Y directions are the same, but the slice thickness (i.e. resolution in the Z-direction) varies. A dashed line indicates isotropy. The exact figures of datasets with resolution greater than or less than 4mm voxel^{-1} in the Z-direction are given in writing above and below the dotted line at 4mm voxel^{-1}.

2.3.2 Random classification forest

We use the voxel-level random classification forest described by Dabbah *et al.* [25]. Briefly, a forest is trained on simple features comprising the Hounsfield Unit values of voxels in the local neighbourhood of each sample voxel. Randomness is achieved by the use of bagging as proposed by Breimen [49], and random subspaces as proposed by Ho [50]. By this we mean that each tree is trained on different, randomly selected, subsets of training data samples and features respectively. The subsets are chosen randomly and uniformly with replacement, such that each tree may be allocated multiple instances of the same *data sample* or *feature*. Landmark detection is then treated as a multinomial classification problem, between the background class c_0 and M landmark classes $c_1 \dots c_M$. During landmark detection, we examine the voxelwise class probability distributions. For each class c_i, $i \neq 0$, the highest-probability voxel in the volume is chosen to be the i^{th} landmark location. The threshold between positive and negative results (i.e. present or not present) is chosen with respect to this classification probability.

The point atlas regression step of [25] is removed and replaced with the atlas location autocontext mechanism of this chapter. Details of the forest classifier are given below. Details of the atlas location autocontext step follow in section 2.3.3.

2.3.2.1 Data pre-processing

Each dataset is pre-processed as follows.

- The slices are concatenated into a volume
- The volume is resampled using linear interpolation to give a volume with slices spaced the same distance apart as the slice pixel resolution, giving isotropic voxels.
- Finally (for efficiency) we run the detection at a low resolution, D_{Res} = 4mm voxel^{-1}. Hence, Gaussian smoothing is applied to avoid aliasing effects and the volume is re-sampled to D_{Res}.

2.3.2.2 Data samples

There are D training datasets. A random subset of D_T training datasets is chosen, uniformly and with replacement, for each tree.

The chance of a dataset being chosen k times may be expressed using the binomial distribution, given the number of trials $n = D_T$ and the chance of a dataset being picked $p = \frac{1}{D}$:

$$B(K = k|n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{2.1}$$

$$B(K = k|D_T, \frac{1}{D}) = \frac{D_T!}{k!(D_T-k)!} \frac{1}{D}^k (1 - \frac{1}{D})^{D_T-k} \tag{2.2}$$

The probability that a dataset is *not* selected for a subset D_T i.e. k = 0, reduces to a simple probability expression.

$$B(K = 0|D_T, \frac{1}{D}) = \left(1 - \frac{1}{D}\right)^{D_T} \tag{2.3}$$

The proportion of datasets selected will then be the complementary probability.

$$\%\ D \text{ datasets in } D_T = 1 - \left(1 - \frac{1}{D}\right)^{D_T} \tag{2.4}$$

In the case that $D_T \approx D$ and where D is large, the Poisson distribution may be used to approximate the binomial distribution. It is easily shown that approximately 37% ($\frac{1}{e}$) of the datasets would not be present in the bootstrap sample.

For each landmark, voxels within a spherical region centred at its position l_i are considered to belong to the landmark class c_i. Samples are given a weighting w, as a Gaussian function of their distance from the landmark position, with a maximum weight of 1.0 at the landmark position. The Gaussian has standard deviation $\sigma_{Sampling}$, and voxels within a radius of $2\sigma_{Sampling}$ are classified as belonging to c_i. All landmark samples in the D_T datasets are used for training.

Remaining data samples are classified as belonging to the background class c_0. A subset of background samples is chosen, uniformly and with replacement, at a ratio B_{Ratio} to the total number of landmark samples. Background samples are assigned a weight $w = 1.0$.

The sample weights are used when computing probability distributions for the purpose of information gain during tree training and for leaf probability distributions, see below.

2.3.2.3 Features

Features are simply the HU intensities of voxels $I(v + d)$ at offsets d from the voxel of interest v. These are sampled from a cuboid neighbourhood with maximum

offset of +/- $\{d_{max}, d_{max}, d_{max}\}$ mm.

Given the volume resolution D_{Res} and the maximum feature offset d_{max} = 52mm, it is straightforward to compute the total number of possible features F.

$$\begin{aligned} F &= \left(2 \times \frac{d_{max}}{D_{Res}} - 1\right)^3 \\ &= \left(2 \times \frac{52}{4} - 1\right)^3 \\ &= 15625 \end{aligned} \tag{2.5}$$

Each tree is provided with a feature subset of size F_T which is randomly selected, uniformly and with replacement.

Additionally, a *sagittal displacement* feature $d_{sag}(v)$ is used, which measures the displacement of a voxel from the mid-volume sagittal plane. The sagittal displacement feature is employed on the basis that most medical datasets are symmetrically aligned about the true mid-sagittal plane of the patient. This feature enables differentiation of, for instance, landmarks on the left (L) and right (R) limbs.

Missing values occur when the feature location is outside of the scan volume. We describe later how to deal with missing values.

2.3.2.4 Forest training

The classification forest consists of a set of T binary decision trees.

Feature selection and information gain

Each binary decision tree is grown by recursively splitting the training data in two, terminating each branch when one of the termination criteria is met (see next section). Starting from the root node, the data S is split into the left and right child node subsets S^L and S^R, according to whether the values for the chosen feature $f \in F_T$ fall above or below the threshold value τ_f. f and τ_f are chosen by running through all combinations of features and thresholds to find the pair which maximises information gain IG.

$$IG = H(S) - \sum_{i \in \{L,R\}} \frac{|S^i|}{|S|} H(S^i) \tag{2.6}$$

H is the Shannon entropy [56], as defined in equation 2.7, over all classes $c = 0 \dots M$.

$$H(S) = -\sum_{c \in C} p(c) log(p(c)) \tag{2.7}$$

$p(c)$ refers to the empirical distribution according to the N weighted samples in the node e.g.

$$p(c) = \frac{\sum_{i=1}^{N} w_i \times 1_c(i)}{\sum_{i=1}^{N} w_i} \tag{2.8}$$

where $1_c(i) = \{1, 0\}$ is an indicator function denoting whether sample i belongs to class c or not.

In fact, following the lead of Quinlan [57], the information gain is computed only over those samples with known values, and the result is scaled by the proportion of values that are known (i.e. not missing). The assumption is that information gain is zero for the samples with missing values.

$$IG = \frac{N_{Known}}{N} \left(H(S_{Known}) - \sum_{i \in \{L,R\}} \frac{|S^i_{Known}|}{|S_{Known}|} H(S^i_{Known}) \right) + \frac{N_{Missing}}{N} \times 0 \tag{2.9}$$

Samples with missing values are sent down both branches of the tree with a half-weighting. This effectively integrates out the nodes for which feature values are not known.

Branch termination criteria

Branches terminate when:

- The remaining samples belong to the same class.
- The total sample weight is less than w_{Node_min}.
- No feature yields information gain.

In theory, we could set a minimum information gain IG_{min}, below which the split is considered trivial. In practice we set IG_{min} to zero but if the smaller of the child nodes contains a sample weight less than $w_{Node\_Split\_min}$, then the split must be trivial and so the branch is terminated.

Leaf probability distributions

End nodes are termed *leaf* nodes. Leaf nodes store estimated class probability distributions which reflect the proportions of training data samples at the node. The total weight of the samples belonging to class c from all N samples which reached a leaf, are defined as follows:

$$w_{Leaf}(c) = \sum_{i=1}^{N} w_i \times 1_c(i) \tag{2.10}$$

The likelihood $L_T(c|\boldsymbol{f_t})$ is the probability that a voxel v belongs to class c given the feature vector $\boldsymbol{f_t}$ formed from $d_{sag}(v)$ plus the F_T features randomly selected from the intensity feature pool $I(v+d)$. It consists of the weight $w_{Leaf\_v}(c)$ of the samples belonging to c in the leaf that v reached, divided by the weight $w_{Tree}(c)$ of the samples for that class at the root node.

$$L(c|\boldsymbol{f_t}) = \frac{w_{Leaf\_v}(c)}{w_{Tree}(c)} \tag{2.11}$$

We subsequently compute posterior probabilities by applying our chosen priors. Equal prior probabilities are assigned to the landmark classes, and a larger prior is used for background (larger by a factor of B_{π_Ratio}). In our scheme, the null class always denotes the background class. Classes 1 to M denote the M landmark classes.

$$\pi(c) := \begin{cases} \frac{B_{\pi\_Ratio}}{B_{\pi\_Ratio}+M} & \text{if } c = 0 \\ \frac{1}{B_{\pi\_Ratio}+M} & \text{otherwise} \end{cases} \tag{2.12}$$

For a novel voxel v, whose feature values are computed from the image data I, the tree posterior probability $P_T(c|\mathbf{f_t})$ of class c is shown in equation 2.13, as a function of the likelihood and the class prior $\pi(c)$.

$$P_T(c|\boldsymbol{f_t}) = \frac{\pi(c)L_T(c|\boldsymbol{f_t})}{\sum_{i=1}^{M} \pi(c_i)L_T(c_i|\boldsymbol{f_t})} \tag{2.13}$$

The forest probability $P_F(c|\boldsymbol{f})$ where $\boldsymbol{f}$ denotes the vector of all features available to the forest (i.e. the union of all $\boldsymbol{f_t}, t = 1 \ldots T$), is computed by taking the mean of the distributions in the leaves of all T trees.

$$P_F(c|\boldsymbol{f}) = \frac{\sum_{t=1}^{T} P_t(c|\boldsymbol{f_t})}{T} \tag{2.14}$$

2.3.2.5 Landmark localisation

The location of each landmark $v^*(c)$ is determined by finding the voxel with the highest probability of belonging to the landmark class, according to the forest posterior probability distribution P_F.

$$v^*(c) = \underset{v \in V}{\arg\max}\ P_F(c|v, I) \tag{2.15}$$

To speed up the algorithm, an initial search is made by testing voxels over a coarse grid of d_{skip} voxel spacing. A local search is then performed on the untested voxels local to the voxel with maximum probability. Finally, Brent interpolation is employed to give a sub-voxel result.

2.3.2.6 Forest shortcut

The *forest shortcut* is a mechanism for reducing the detection run time. The principle is that many voxels can be confidently classified as background after evaluating far fewer than T trees. For every tree after some minimum number T_{min} have been evaluated, we run a statistical test on the accumulated results. If all landmarks are found to have an estimated probability $P_F(c|\boldsymbol{f})$ less than some threshold $P_{Shortcut}$ within a certain confidence, then we abort evaluation and report the current class probabilities.

The statistical test is conducted as follows. We assume that the individual trees give independent probability predictions, with approximately equal accuracy. For any given voxel, the individual tree predictions $P_T(c|\boldsymbol{f_t})$ for each class c are assumed to follow a normal distribution centred at some intrinsic prediction probability $P_{T\_Intrinsic}(c)$. The more trees that have been evaluated, the more confident we can be that the mean $P_T(c|\boldsymbol{f_t})$ (= $P_F(c|\boldsymbol{f})$) approximates $P_{T\_Intrinsic}(c)$. We compute the standard error of $P_T(c|\boldsymbol{f_t})$, and if $P_{Shortcut}$ lies at least three standard errors above, then we assume that $P_{T\_Intrinsic}(c)$ lies below $P_{Shortcut}$.

So, if for all landmark classes $c_1 \ldots c_M$,

$$\mu(P_T(c_i|\boldsymbol{f_t})) + 3\frac{\sigma(P_T(c_i|\boldsymbol{f_t}))}{\sqrt{T_{Evaluated}}} < P_{Shortcut} \tag{2.16}$$

where μ and σ are the mean and standard deviation of $P_T(c_i|\boldsymbol{f_t})$ over all trees $t = 1 \ldots T_{Evaluated}$, then the voxel is assumed to be background, and we do not evaluate any further trees for the voxel in question.

The assumption of a normal distribution is an approximation. For comparison, we direct the reader to a similar method proposed by Schwing *et al.* [58] which sets out a theoretical justification for the beta distribution.

2.3.2.7 Parameter values

Dabbah *et al.* [25] chose parameter values (see Table 2.1) empirically from cross-validation experiments. The classifier is relatively insensitive to changes in the current parameter settings. Further small gains in accuracy (but large decreases in speed) might be achieved by increasing the number of datasets per tree, or the number of trees. For other parameters, the optimum values are already selected in terms of accuracy.

In this chapter, we do not attempt to re-explore the parameter space, but focus on the addition of atlas location autocontext and its effect on a pre-existing classifier. Later, different parameter settings are explored for specific applications. In particular, see section 3.4 for an illustration of parameter tuning in the case of gradient orientation features.

Parameter	Definition	Value
D_{Res}	Resolution at which detector is run	4mm voxel^{-1}
T	Number of trees in forest	80
D	Number of training datasets	272
D_T	Number of training datasets sampled per tree (bagging)	40
d_{max}	Maximum feature offset	52mm
F	Total number of possible features	15,625
F_T	Number of features selected per tree	2500 + 1 ($d_{sag}(v)$)
$\sigma_{Sampling}$	Standard deviation of Gaussian weighting function for landmark samples	3.0mm
B_{Ratio}	Ratio of background to foreground training samples	5.0
B_{π_Ratio}	Ratio of background class to landmark class prior probability	400
w_{Node_min}	Minimum total weight of samples in a node for branch splitting, otherwise branch is terminated.	5.0
$w_{Node\_Split\_min}$	Minimum total weight of samples in smallest child node, otherwise branch is terminated.	2.0
d_{skip}	(Detection phase) Grid search interval	2 voxels
T_{min}	(Detection phase) Minimum number of trees to evaluate before forest shortcut may be deployed.	5 trees
$P_{Shortcut}$	(Detection phase) Minimum probability for forest shortcut.	0.15

Table 2.1: Empirically chosen parameter values for the random classification forest for anatomical landmark detection in whole-body CT.

2.3.3 Atlas location autocontext

In this section we describe the atlas location autocontext mechanism that we are proposing, and we present experiments performed with the training data to choose the best atlas mapping method.

2.3.3.1 Feedback of atlas location features

The landmark random forest detector is run exactly as described in section 2.3.2. We term this the *zeroth* detector $P_F(c|\boldsymbol{f})^0$. A mapping is then computed between the detected landmark locations and a landmark atlas (i.e. from *volume* space to *atlas* space). We constructed the atlas by affinely registering all the training ground truth points to the ground truth of an arbitrarily chosen reference whole-body dataset (using least squares fitting) and then taking the mean position of each landmark.

There are a number of options for mapping from volume space to atlas space. The mapping takes the form $T_a(v) : \mathbb{R}^3_{vol} \mapsto \mathbb{R}^3_{atlas}$. In section 2.3.3.2 we describe mapping variants which range from a global transform with constraints on scale and skew to a fully flexible non-rigid spline mapping. Section 2.3.3.3 contains an evaluation of the accuracy of these mappings.

After finding the mapping $T_a(v)^0$, we then train a new detector $P_F(c|\boldsymbol{f})^1$, in exactly the same way as $P_F(c|\boldsymbol{f})^0$ except that we give the x, y and z components of the estimated atlas coordinate $T_a(v)^0$ as scalar features to each tree of the forest in addition to the original features. The feature vector $\boldsymbol{f}$ now contains a mix of features: the sagittal displacement feature $d_{sag}(v)$, the intensity features $I(v+d)$ and the atlas coordinates $T_a(v)$. We continue to iterate, training detectors $P_F(c|\boldsymbol{f})^1, P_F(c|\boldsymbol{f})^2, P_F(c|\boldsymbol{f})^3 \ldots P_F(c|\boldsymbol{f})^n$ on the atlas coordinate features $T_a(v)^0, T_a(v)^1, T_a(v)^2 \ldots T_a(v)^{n-1}$, in an ongoing feedback loop.

In the rare case where no landmarks are detected, the atlas feature values are treated as missing and dealt with in the same way as missing image intensity values. For these datasets, the feedback loop is redundant and we would not expect the results to change with iterations beyond the zeroth.

2.3.3.2 Description of different mappings to atlas space

A summary of the mappings that we consider is given in Table 2.2. These map from the set of N landmarks $l_1 \ldots l_N$ with detection probability $P_F(c|\boldsymbol{f}) \geq \tau_P$ to their corresponding point atlas locations $a_0, a_1 \ldots a_N$. The threshold τ_P is applied

Mapping	Global vs. Local	Degrees of freedom
Similarity	Global	7
Affine	Global	12
Gaussian spline (Affine)	Local	$(3 \times N) + 1$
Gaussian spline (Similarity)	Local	$(3 \times N) + 1$
Thin plate spline (Affine)	Local	$3 \times N$
Thin plate spline (Similarity)	Local	$3 \times N$

Table 2.2: Summary of six different spatial mapping types. The splines have 3 degrees of freedom, corresponding to three spatial dimensions, for each of the N landmarks. The Gaussian spline has an extra parameter σ specifying the standard deviation of the Gaussian function.

to reduce the number of false positives. Details of how to choose τ_P are given in section 2.3.3.3.

Descriptions of the mappings follow. A visual illustration is shown in Figure 2.7 for a thoracic dataset. It can be seen that there is variation in the orientation and shape of the mappings.

Global transforms

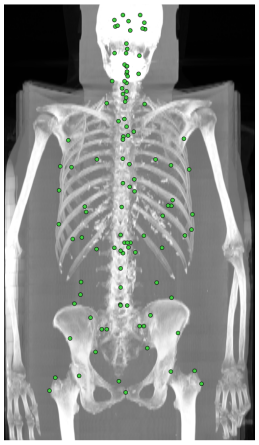
The similarity transform mapping is composed of a 3×3 transformation matrix A and a translation vector t, but A is constrained to allow isotropic scaling and rotation only. There are thus seven degrees of freedom comprising global translation (in the x, y and z directions), global rotation (around the x, y and z axes) and a uniform scaling factor applied to all three dimensions.

$$T_a(v) = Av + t \tag{2.17}$$

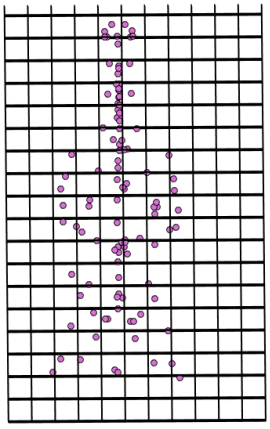
A and t are found by least squares fitting.

$$T_a(v) = \underset{T_a(v)}{\arg\min} \sum_{i=1}^{N} |l_i - T_a(l_i)|^2 \tag{2.18}$$

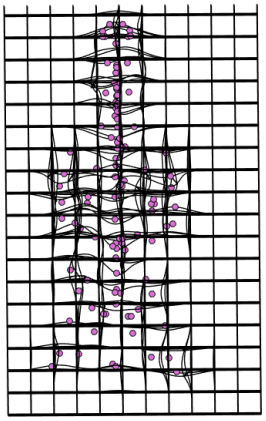
An affine transform has the same formulation as the similarity transform. However, all twelve degrees of freedom are permitted (i.e. A is not constrained), comprising three-directional translation, rotation, scaling and skew.



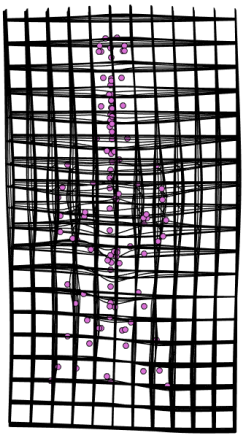
(a) Thorax dataset



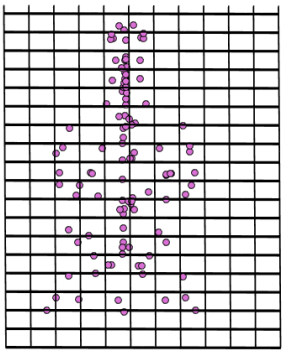
(b) Similarity (sim) transform



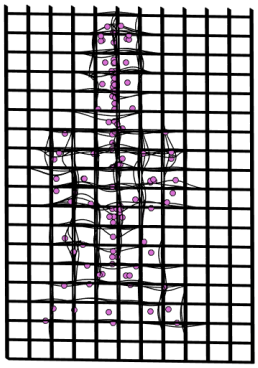
(c) Gaussian sim. spline



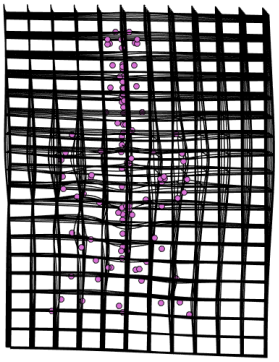
(d) Thin plate sim. spline



(e) Affine (aff) transform



(f) Gaussian aff. spline



(g) Thin plate aff. spline

Figure 2.7: Visual illustration of six different mappings to atlas space, from the ground truth landmarks to the landmark atlas, for an example thorax dataset. The images show a grid with 20mm spacing which has been mapped from volume space to atlas space. The viewpoint for each is chosen to best view the shape of the grid. This is seen in the slightly differing configurations of the landmarks.

Local spline transforms

For the purpose of non-rigid mapping, we consider splines built as the sum of a set of radial basis functions $\phi(v, l_i)$. A radial basis function is a real-valued function whose value depends only on the distance between v and a centre point which in this case would be the i^{th} landmark point l_i, so e.g. $\phi(v, l_i)$ can also be expressed as $\phi(r)$ where $r = |v - l_i|$. The spline takes the form

$$T_{a\_dim}(v) = p(v) + \sum_{i=1}^{N} w_i \phi(|v - l_i|) \tag{2.19}$$

$p(v)$ is a polynomial $a_0 + a_1 v_x + a_2 v_y + a_3 v_z$ specifying the underlying global affine transformation as a linear function of the x, y and z components of v. The weights $w_i = w_1 \dots w_N$ are associated with each of the N radial basis functions, whose centres correspond to landmarks $l_i = l_1 \dots l_N$.

One spline mapping function is required for each of the three Cartesian coordinate dimensions $c \in \{x, y, z\}$ so the atlas transform $T_a(v)$ comprises $\{T_{a\_x}(v), T_{a\_y}(v), T_{a\_z}(v)\}$. The polynomial coefficients and spline weights are computed using the method in [59], such that $T_a(l_i)$ is equal to a_i. In other words, all detected landmarks map exactly to the corresponding atlas landmarks. In general, $d + 1$ detected landmarks are required to solve the spline equations, where d is the number of dimensions, and they should span the full d dimensions i.e. not be collinear (2D) or coplanar (3D). Thus in three dimensions, at least four detected landmarks are required. In the case of exactly four landmarks, the solution reduces to the least-squares regression affine solution, which is fully described by $p(v)$.

For the thin plate spline, the RBF is simply $\phi(r) = r$. This is the fundamental solution of the biharmonic equation [59, 60] (which is the Laplacian operator squared) i.e. if we find a function that has a fourth derivative equal to zero except at the landmarks, then the solution produces a smooth deformation between points which minimises a type of "bending energy". So, where δ is the Dirac delta function:

$$\nabla^4 T_a(v) = \Delta^2 T_a(v) = \sum_{i=1}^{N} w_i \, \delta(|v - l_i|) \tag{2.20}$$

The general solution is (with the addition of the affine part $p(v)$ to regularise the behaviour of $T_a(v)$ at infinity):

$$T_a(v) = p(v) + \sum_{i=1}^{N} w_i \, \phi(|v - l_i|) \tag{2.21}$$

which in the case of three dimensions reduces to:

$$T_a(v) = p(v) + \sum_{i=1}^{N} w_i \, |v - l_i| \tag{2.22}$$

For the Gaussian spline, $\phi(r) = e^{-\frac{1}{2}\left(\frac{r}{\sigma}\right)^2}$ where $r \geq 0$. In this spline, each landmark influences only the local mapping (within approximately 3 σ). The choice of σ thus has an appreciable affect on the shape of the resulting warp. Empirical testing showed the most accurate results were achieved with $\sigma = 30$mm. Accuracy was measured in terms of the mean landmark error, see section 2.3.3.3.

We trial a modification to the splines in which the underlying affine transformation $p(v)$ is constrained to be the similarity transform described above. The idea is that this will prevent anatomically unrealistic affine transformations with large degrees of skew or highly non-isotropic scaling, as sometimes happens when the number of detected landmarks is small or when they lie in similar planes.

2.3.3.3 Evaluation of mapping accuracy

We assess the accuracy of the six mappings described in section 2.3.3.2. We expect a non-rigid mapping to better model inter-subject anatomical variation (as described in 1.2) than a global transform. For one thing, body proportions and shapes vary widely subject to subject. For another, patient posture may make a large difference as articulating body parts can be in significantly different positions relative to one another i.e. knees bent or straightened, head tucked into chest or upright, arms up or down by the sides. Hence a single global transformation is likely to give a poor mapping between atlas and subject.

However, the flexibility of the non-rigid spline mappings mean that falsely detected landmarks may cause greater distortion. Hence our inclusion of the Gaussian spline alongside the more standard thin plate spline, because the impact of any given landmark (including false positives) is limited to a local sphere of influence.

Method

A random classification forest is trained on all training datasets, as described

in section 2.3.2. Classification is run on the training data cohort and the detected landmarks are extracted. We then threshold the landmarks at $P_F(c|\boldsymbol{f}) \geq \tau_P$ where $0 \leq \tau_P \leq 1.0$, and construct a mapping $T_v(v)$ from the landmark atlas to these landmarks. Note that we are measuring the errors in volume space rather than atlas space i.e. we map in the opposite direction to that required for the atlas coordinate features. This is simply for ease of comparison with results from the main method (Figure 2.10 later shows how direct correction by atlas registration compares with atlas location autocontext).

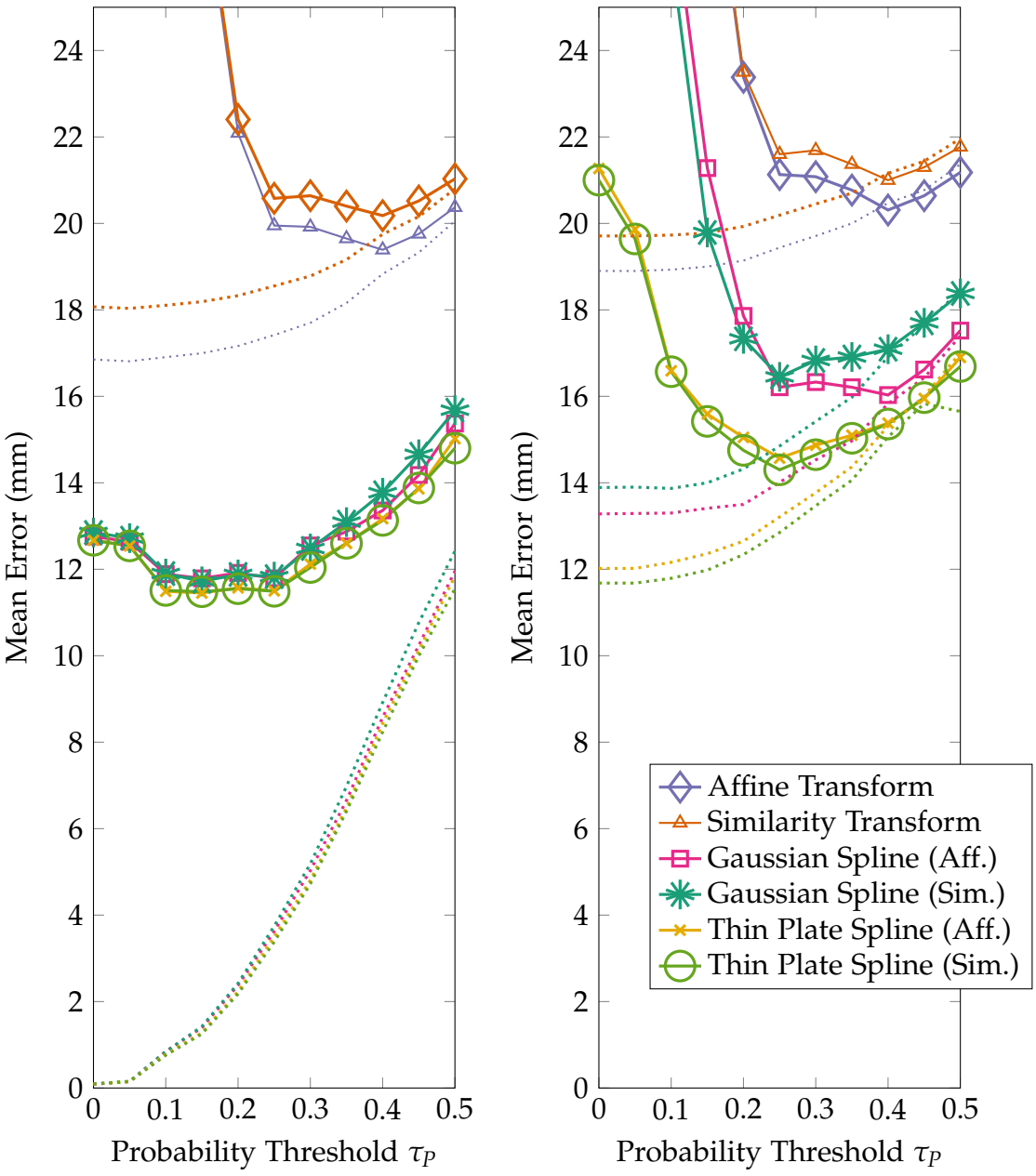
We employ a conservative strategy for cases where there are fewer than six landmarks, in order to prevent overfitting to the data. A similarity transformation is used for the case of five landmarks. A simple best fit translation is used for four landmarks or fewer.

Results

A graph of the results is shown in Figure 2.8. Assessment of the accuracy of each mapping type is done by measurement of the mean landmark error and mean inter-landmark error on the 272 training datasets. The *landmark* error is measured between each landmark and its mapped atlas equivalent when all landmarks $P_F(c|\boldsymbol{f}) \geq \tau_P$ are included in the mapping computation. The *inter-landmark* error is estimated by excluding the landmark for which the error is being measured from the mapping computation; in the case of false negative landmarks $P_F(c|\boldsymbol{f}) < \tau_P$, the landmark and inter-landmark errors are the same.

To give further insight, and as a sanity check, the procedure is repeated using the ground truth landmarks in place of the detected landmarks. At each threshold, a mapping is made from the ground truth landmarks which *correspond* to those detected landmarks with $P_F(c|\boldsymbol{f}) \geq \tau_P$. The ground truth results are shown with dotted lines on the graph.

Some results are omitted from the graph. One excluded dataset gives very large mapping errors (100-600mm) because of poor detection results. At values of τ_P greater than or equal to 0.4, there are datasets with no positively detected landmarks which are also excluded (5, 7 and 11 datasets at τ_P = 0.4, 0.45 and 0.5 respectively). Hence the results look better than they are in actuality, however since the number of datasets is small and since the error is already worsening for probabilities of 0.4 and higher, it is assumed that the trend remains valid although we get a deceptively more modest decline.



(a) Landmark error

(b) Inter-Landmark Error

Figure 2.8: Graphs comparing the registration accuracy of six different mapping methods to atlas space, evaluated on the training data. The solid lines show the results using the detected landmarks (after the zeroth iteration), and the dotted lines show the results using the ground truth landmarks. The splines are shown with affine (Aff.) and similarity (Sim.) underlying global transforms. a) The *landmark* error is measured between each landmark and its mapped atlas equivalent when *all* landmarks are included in the mapping computation. b) The *inter-landmark* error is estimated by excluding the landmark for which the error is being measured from the mapping computation.

Ground Truth Landmarks To Atlas: These results are shown as a control, to explore the accuracy of mappings in the case of perfect landmark detection. As expected, the splines have perfect landmark accuracy when all landmarks are included, at $\tau_P = 0$. The global transforms perform significantly worse. As the number of landmarks is reduced, the accuracy decreases. The thin plate spline with underlying similarity transform is marginally more accurate than the other spline variants in terms of landmark error. In terms of inter-landmark error, the thin plate splines give significantly better accuracy than the Gaussian splines.

Detected Landmarks To Atlas: The ranking of the mappings in order of accuracy is the same for detected landmarks as for the ground truth. The mappings initially become more accurate as τ_P increases, because false positive results are removed. This is particularly obvious through inspection of the spline inter-landmark errors. As τ_P increases further, the accuracy decreases because of the omission of accurately detected landmarks. Notice that the spline landmark error at $\tau_P = 0$ is exactly equal to the mean landmark error of the random forest detector.

Conclusion

According to both mean landmark and mean inter-landmark error, the thin plate spline has best performance. If the spline's underlying global transform is constrained to be a similarity transform, this appears to give a marginal improvement. The Gaussian spline fails because the finite support of the landmarks lead to a non-smooth mapping (as illustrated in Figure 2.7) which is less biologically cogent.

In the remainder of the chapter we show results for landmark detection experiments using both the thin plate similarity spline and the affine transform, which are the best performing local and global transforms respectively. These two mappings behave differently and should provide an interesting comparison.

In section 2.3.3.4 we look at how to identify and remove false positive results from the mappings to give improved accuracy.

2.3.3.4 Elimination of false positives from the atlas mapping

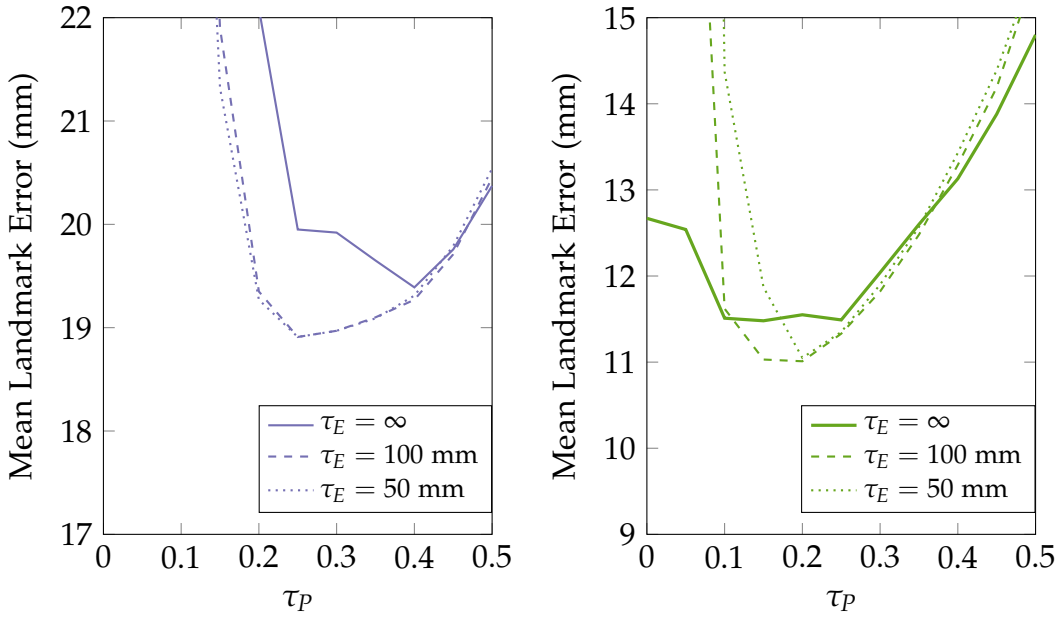
We employ a method of iterative fitting to improve the accuracy of the mapping using the detected landmarks. In summary, landmarks are iteratively removed from the mapping, which is subsequently recomputed, until the mapping fit error is less than a specified distance τ_E for all landmarks in the mapping.

For the affine transform, the mapping error is defined as the distance between the atlas landmark and the transformed detected landmark i.e. landmark error. However for the thin plate spline, all positive detected landmarks are exactly mapped to the atlas landmarks. Hence we use the inter-landmark error for the purpose of spline mapping rejection. We note that for only 127 landmarks, a leave-one-out method (as required by the inter-landmark error) is feasible; for a great number of landmarks, this might become rather less so. Algorithm 1 gives a more formal definition of this process.

Algorithm 1 Atlas mapping by Iterative fitting
Algorithm for mapping the set of detected landmarks L_D to the set of atlas landmarks L_A, given the probability threshold τ_P and the fit error threshold τ_E. Two trimmed subsets are discovered, L_{ATr} and L_{DTr}, for which all fit errors fall below τ_E when one is mapped to the other. This mapping $T_A(v)$ is returned.

procedure TRIMMEDFIT(L_D, L_A, τ_P, τ_E)
  $L_{DTr} \leftarrow L_D \geq \tau_P$ ▷ Choose detected landmark set with $P_F(c|\boldsymbol{f}) \geq \tau_P$
  $L_{ATr} \leftarrow L_A \in L_{DTr}$ ▷ Choose corresponding atlas landmarks
  $E_{max} \leftarrow \infty$
  while $E_{max} > \tau_E$ **do** ▷ We have the answer if $E_{max} \leq \tau_E$
    $E_{max} \leftarrow \tau_E$
    $l_{max} \leftarrow NULL$
    $T_A(v) \leftarrow$Fit(L_{DTr}, L_{ATr}) ▷ Find mapping between trimmed sets
    if Affine Fit **then**
      $T_{AE}(v) = T_A(v)$
    end if
    for all $l_i \in L_{DTr}$ **do** ▷ Find maximum error.
      if Spline Fit **then** ▷ Spline case: Leave-one-out mapping.
        $T_{AE}(v) \leftarrow$SplineFit$(L_{DTr} \neq l_i, L_{ATr} \neq l_i)$
      end if
      $E_i \leftarrow |T_{AE}(l_i) - (l_i \in L_{ATr})|$
      if $E_i > E_{max}$ **then**
        $E_{max} \leftarrow E_i$
        $l_{max} \leftarrow l_i$
      end if
    end for
    if $l_{max} \neq NULL$ **then**
      $L_{DTr} \leftarrow L_{DTr} \neq l_{max}$ ▷ Remove landmark with maximum error.
      $L_{ATr} \leftarrow L_{ATr} \neq l_{max}$
    end if
  end while
  return $T_A(v)$
end procedure

There are now two parameters to set when computing the atlas map-



(a) Affine Transform

(b) Thin Plate Similarity Spline

Figure 2.9: Graphs showing the interplay between the τ_P and τ_E iterative fitting thresholds for mapping detected landmarks to atlas space.

ping: the forest probability threshold τ_P and the distance error threshold τ_E. During training of the algorithm, parameter values are chosen by running through the training results and finding the pair of values which give the lowest mean landmark error. We perform grid search over all combinations of $\tau_p \in \{0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6\}$ and $\tau_E = \{30, 50, 75, 100, 150, \infty\}$. We have observed empirically that the shape of this two-dimensional space is convex, so we assume that this simple method suffices to find a stable operating point.

An illustration is shown in Figure 2.9 of the values obtained for different rejection error thresholds for each of the two mappings. The plots at a rejection distance of infinity are equivalent to the plots shown in Figure 2.8.

In each feedback iteration, thresholds τ_P and τ_E are learnt automatically by comparison of the training data with its ground truth. We store these threshold values $\tau_{P\_0}, \tau_{P\_1}, \tau_{P\_2} \ldots \tau_{P\_n-1}$ and $\tau_{E\_0}, \tau_{E\_1}, \tau_{E\_2} \ldots \tau_{E\_n-1}$ along with the detectors $P_F(c)^1, P_F(c)^2, P_F(c)^3 \ldots P_F(c)^n$, for later application to the test datasets.

2.3.4 Out-of-bag detection and resubstitution during forest training

Out-of-bag detection

During the training phase, we run detection on the training datasets in order to obtain atlas coordinate features, and to train the atlas mapping parameters. An out-of-bag strategy is employed where each dataset is classified only by the subset $T_{Out-of-bag}$ of trees in the forest which have not been trained on that dataset. This is made possible since each tree holds a list of the ID numbers of the datasets on which it was trained.

The expected size of the forest subset $T_{Out-of-bag}$ for any given training dataset is a little smaller than T. We multiply the chance that a dataset is not selected for a tree (refer back to equation 2.3) by the number of trees T to arrive at equation 2.23.

$$T_{Out-of-bag} = T \times \left(\frac{D-1}{D}\right)^{D_T} \tag{2.23}$$

For $D = 272$, $D_T = 40$ and $T = 80$, the expected value of $T_{Out-of-bag}$ is 69.0 trees.

Test datasets are reserved at the start and kept separate during the training process. The size of the forest used in the test phase is simply T i.e. $T_{Out-of-bag} = T = 80$.

Resubstitution

To completely avoid resubstitution during training of the forest, a fresh set of data would be required to train each iteration. Given the limited number of training datasets $D = 272$, it was decided to use all training data in each pass.

As a result, there is a weak element of resubstitution from pass 1 onwards because each tree is given access not only to the $D_T = 40$ training datasets but also to the 40 corresponding mappings which have been derived from the forest results, and hence contain information from many ($\rightarrow D$) datasets which are not explicitly considered to be training datasets for the tree. We hypothesise that the contribution to the mapping from any individual dataset is small enough so as to be negligible. Inspection of the training data (see Figures 2.10 and 2.11) results shows that the training and test results exhibit a similar pattern, suggesting that

resubstitution has not caused overfitting to the training data.

Again, the independent test dataset population is reserved at the start and kept separate during the training process. There is obviously no resubstitution in the test results.

2.4 Evaluation

2.4.1 Detection and localisation results

We trained and validated the algorithm for 7 iterations, yielding a zeroth detector equivalent to the original detection algorithm, and six subsequent detectors to which results from the previous detector were fed to give atlas coordinate features.

Results for mean error and AUC are shown in Figures 2.10 and 2.11 respectively. Test and training results follow a similar pattern, suggesting that overfitting due to resubstitution is not occurring.

2.4.1.1 Mean Error

There is an obvious improvement in the results between the zeroth and first iterations, from a mean of 12.49mm to 9.96mm (thin plate spline) and to 9.86mm (affine transform). There is further significant improvement after the second iteration to 9.57mm (thin plate spline) and to 9.52mm (affine transform). Over the course of further iterations, there is some noise in the results which is similar to the level of random noise in the process.

The black dotted lines indicate the mean error if spline interpolation was used after iteration zero to correct those landmarks with forest probability $P_F(c|\boldsymbol{f}) < \tau_P$ i.e. directly inferring landmark locations from the spline mapping. It can be seen that direct interpolation gives a smaller improvement than using the atlas coordinate feature, which vindicates the usage of a(nother) machine learning step to exploit this information.

2.4.1.2 Area Under the Curve

A *localisation* Receiver Operating Curve (ROC) curve is used in order to express the detection performance for an acceptable location error, in this case to within 30mm of the ground truth. There is significant improvement between the zeroth and second iterations, which is better in the case of the affine transform (0.893 to 0.930) than that of the thin plate spline (0.893 to 0.919). Again, in subsequent iterations there is no significant change.

On balance, one iteration appears to confer the majority of the improvement. Two iterations is optimum, particularly in terms of the AUC results.

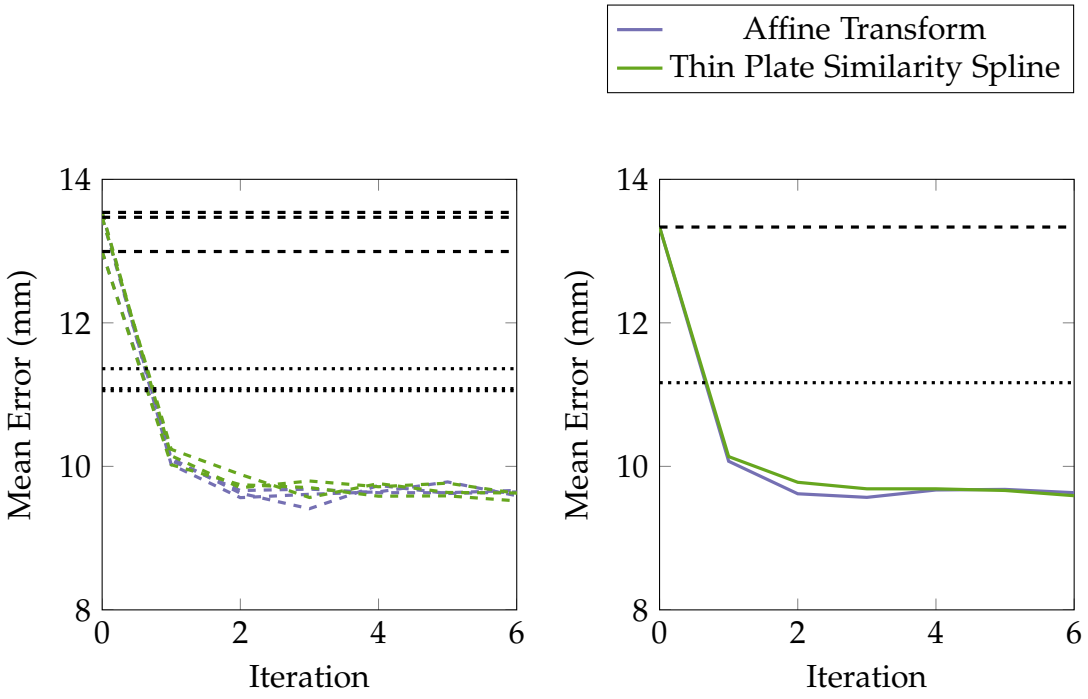
Significance tests for the above results were performed using a two-tailed

paired Student's t-test ($p < 0.001$). Associated F-tests were performed to check that the variance in the two populations under comparison were not significantly different.

2.4.2 Run times

The test computer has a dual-processor, 24-core Intel Xeon CPU with clock speeds of 2.40GHx and 2.39GHz, and 32GB RAM.

Training takes approximately 11 hours for the zeroth detector, and 9 hours for each detector thereafter, since the trees are smaller for the first and following detectors. The mean detection time for a novel dataset is approximately 2.5 seconds for the zeroth detector and 1.5 seconds thereafter. These times are viable for clinical usage.



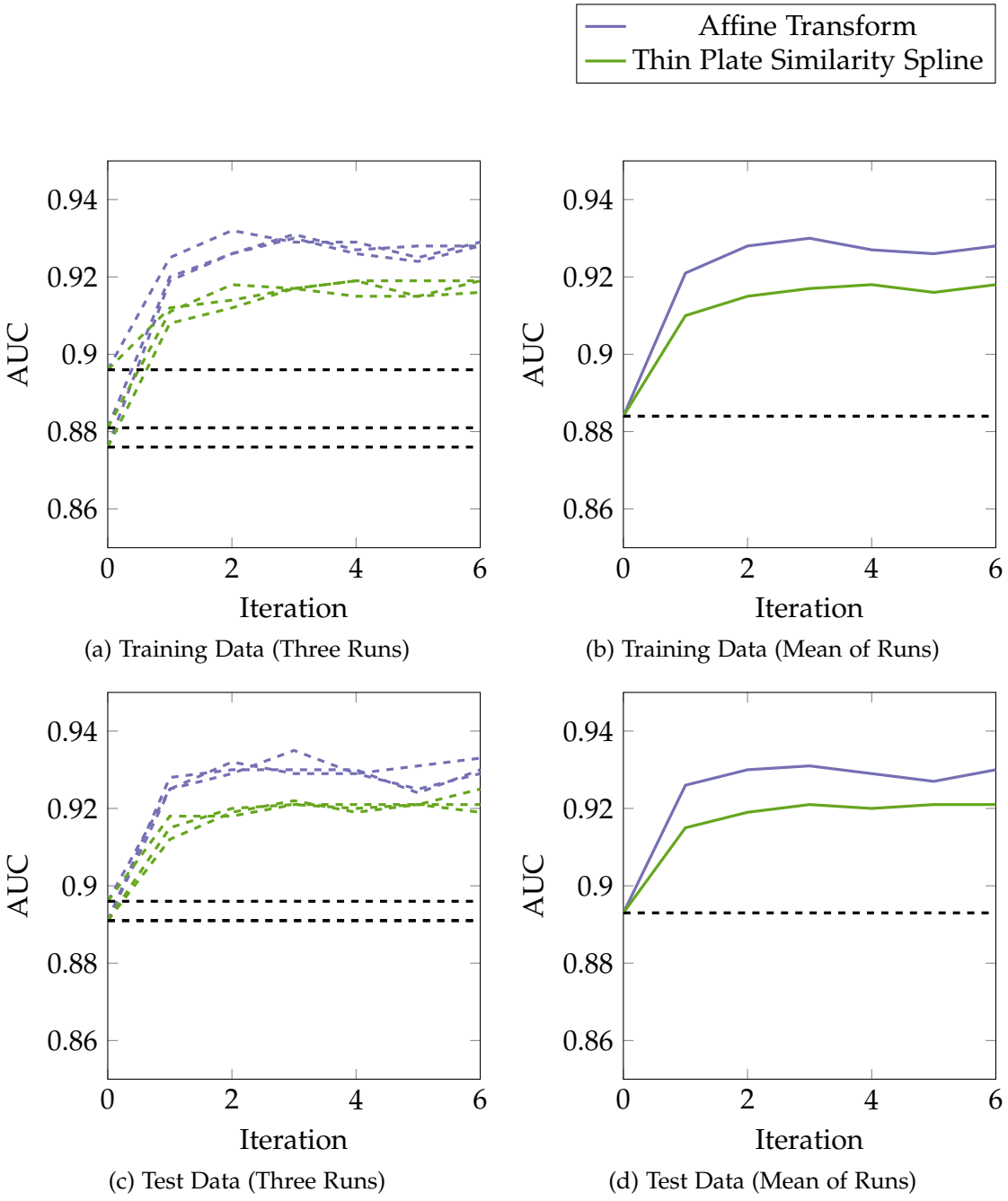
(a) Training Data (Three runs)

(b) Training Data (Mean of Runs)

(c) Test Data (Three Runs)

(d) Test Data (Mean of Runs)

Figure 2.10: Graphs showing the mean error over the course of seven iterations. The results of three separate experiment runs are shown with dashed lines (left-hand graph) and the overall mean result is shown with a solid line (right-hand graph). The black dashed lines indicate the mean error at iteration 0. Iteration 0 is the baseline result without atlas location autocontext. The black dotted lines indicate the mean error if spline interpolation was used after iteration zero to correct those landmarks with forest probability $P_F(c|\boldsymbol{f}) < \tau_P$.



(a) Training Data (Three Runs)

(b) Training Data (Mean of Runs)

(c) Test Data (Three Runs)

(d) Test Data (Mean of Runs)

Figure 2.11: Graphs showing the AUC over the course of six iterations. The results of three separate experiment runs are shown with dashed lines (left-hand graph) and the overall mean result is shown with a solid line (right-hand graph). The black dashed lines indicate the AUC at iteration 0. Iteration 0 is the baseline result without atlas location autocontext.

2.4.3 Analysis of the detected landmarks

A high-level evaluation of the results demonstrates that atlas location autocontext gives a significant, if modest, improvement in the mean error and detection rate. Taking one run and looking more closely at the results, we can now see where the improvement is coming from. We will compare the zeroth iteration with the second iteration.

2.4.3.1 Landmark performance by body compartment

Tables 2.3 and 2.4 show the landmark performance broken down by body compartment. Mean AUC improves for all body compartments following feedback. Mean error improves dramatically in the thorax and lower limb following feedback, and a little less dramatically in the abdomen. However, accuracy in the head is already good in the zeroth pass, and worsens slightly with atlas location feedback. We propose in section 2.6.1.2 that a locally affine mapping might give better performance.

Body Compartment	**Pass 0**	**Pass 2 (Affine)**	**Pass 2 (Spline)**
Head and Neck	6.03	6.41	7.32
Thorax	21.34	10.82	11.30
Abdomen	14.87	12.00	12.09
Lower Limbs	18.36	7.54	7.25
All	12.49	9.52	9.57

Table 2.3: Mean landmark errors (mm), broken down by body compartment.

Body Compartment	**Pass 0**	**Pass 2 (Affine)**	**Pass 2 (Spline)**
Head and Neck	0.987	0.991	0.987
Thorax	0.883	0.932	0.903
Abdomen	0.892	0.923	0.909
Lower Limbs	0.953	0.998	0.997
All	0.893	0.930	0.919

Table 2.4: Mean AUC, broken down by body compartment.

2.4.3.2 Individual landmark errors

The change in mean error *per landmark* is shown in Figure 2.12. The results improve for all landmarks following feedback. The improvement is most obvious in the case of landmarks which are poorly detected in the zeroth iteration. The change is much less dramatic in the head and neck, where detection results are consistently good from iteration zero onwards.

2.4.3.3 Images of the detected landmarks

Maximum Intensity Projection (MIP) images of the results are presented for a number of example datasets in Figures 2.13 to 2.16. A probability threshold of 0.2 is used to distinguish positive from negative detection results. The symbols used in the images are as follows:

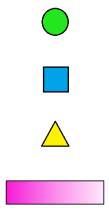
●	Ground truth landmark position
■	Positively detected landmark position, $P_F(c\|\boldsymbol{f}) \geq 0.2$
▲	Negatively detected landmark position, $P_F(c\|\boldsymbol{f}) < 0.2$
✖	Falsely detected landmark position, $P_F(c\|\boldsymbol{f}) \geq 0.2$
●—■	Link between corresponding ground truth and detected positions

We can see at a glance that outliers are moved to more sensible positions. It is apparent that atlas location autocontext aids in better spatially placing the "uncertain" landmarks which were described in section 2.3.1, which is a satisfactory outcome. In particular, see the vertebra in Figure 2.14 (and probability cloud in Figure 2.20) and the pancreas head in Figure 2.16 (and probability cloud in Figure 2.21). We did not include uncertain landmarks in the numerical results, so this is a "silent" improvement.

2.4.3.4 Examining the landmark probability clouds

The mechanism of atlas location autocontext is illustrated in Figures 2.17 to 2.24. A pink overlay shows the voxelwise probabilities $P_F(c|\boldsymbol{f})$ for each given landmark class. The intensity of the pink colour corresponds to the magnitude of the probability. The key to these images is as follows:

Generally those voxels closest to the landmark of interest have highest probability, causing a pink cluster or "cloud" close to the true position, with other clouds congregating around regions of similar appearance elsewhere in the vol-



Ground truth landmark position

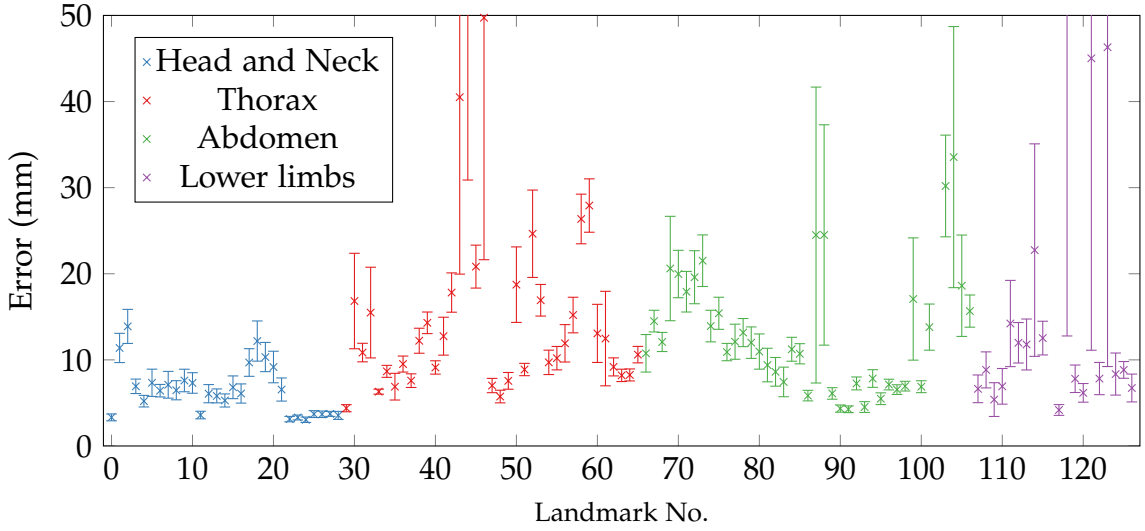
Positively detected landmark position, $P_F(c|\boldsymbol{f}) \geq 0.2$

Negatively detected landmark position, $P_F(c|\boldsymbol{f}) < 0.2$

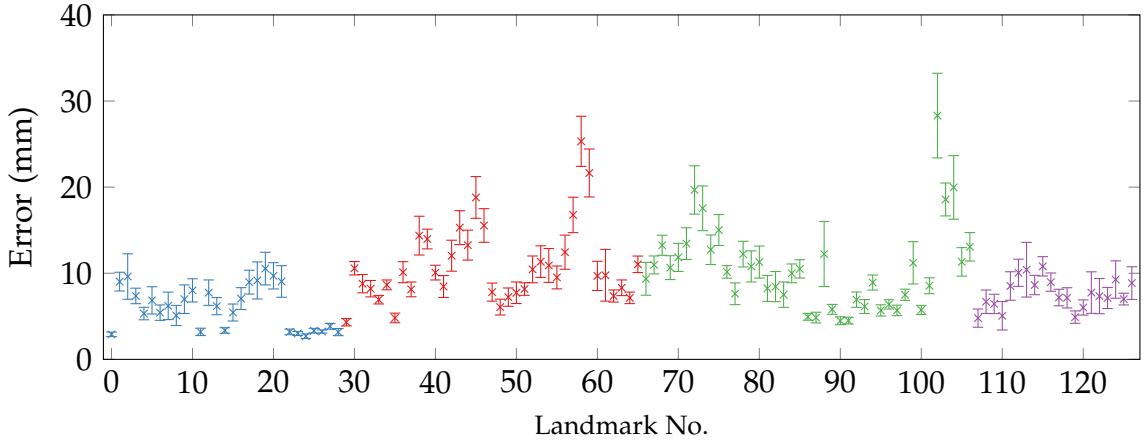
Probability shading, (L $\rightarrow$ R, maximum $\rightarrow$ minimum probability)

ume. The clouds are more diffuse for landmarks which are not present or are not easily visible in the scan.

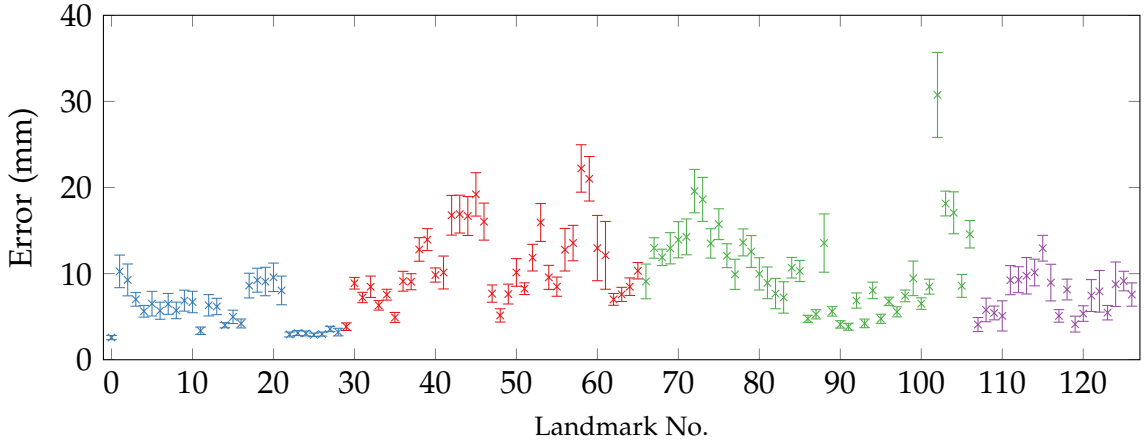
It can be seen that for each landmark, the probability cloud is better localised in iteration two compared to iteration zero. The probability clouds have very similar shape for both choices of mapping. This suggests that the forest is reaping similar spatial information from both choices of mapping. It appears that neither is accurate enough to solve problems such as that of rib and vertebrae repeating structures (see Figure 2.13).



(a) Iteration 0

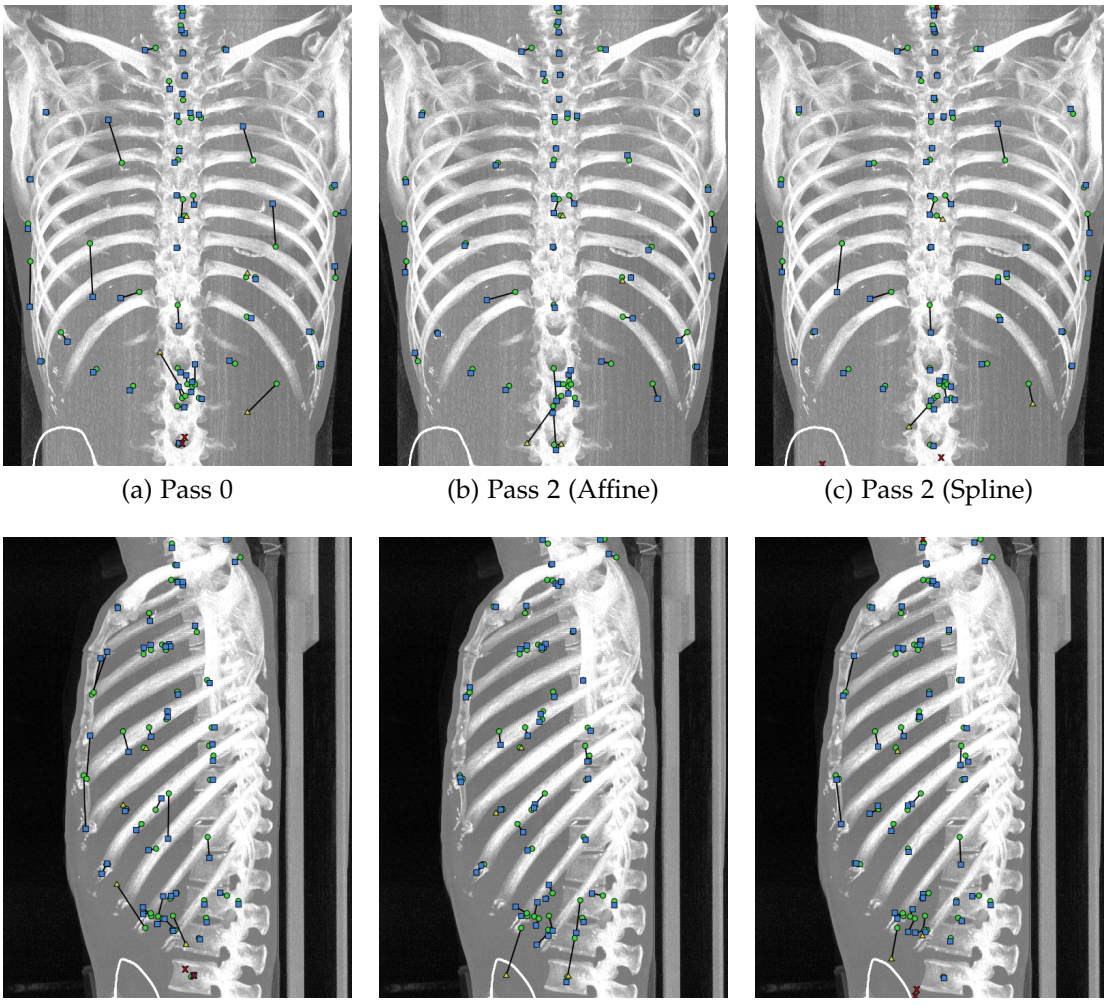


(b) Iteration 2 (Affine transform)



(c) Iteration 2 (Thin Plate Similarity Spline)

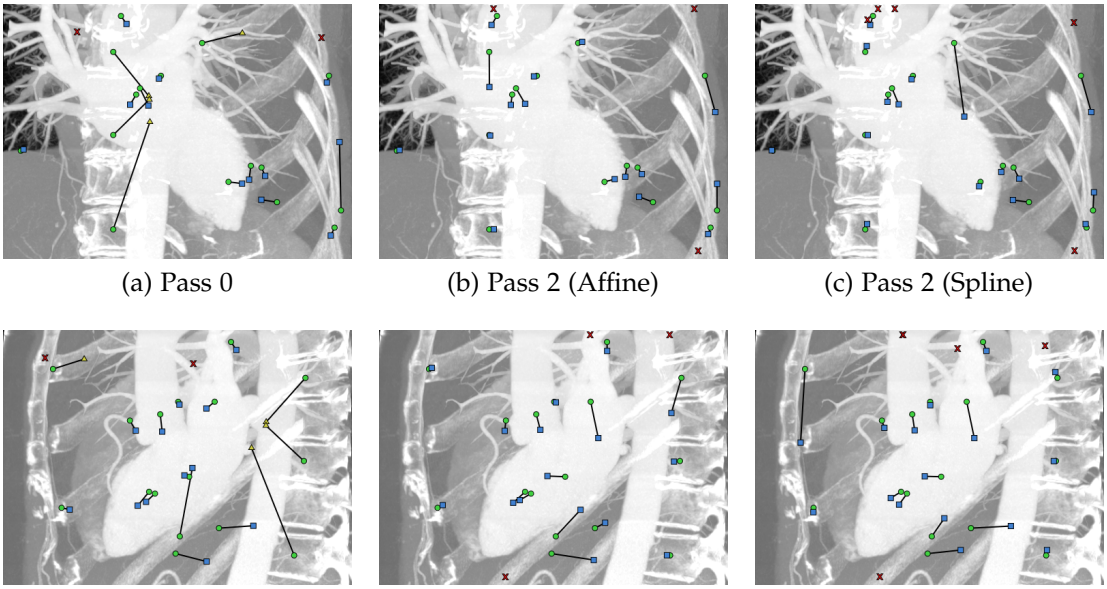
Figure 2.12: Error bars for individual landmarks (mean error +/- standard deviation).



(a) Pass 0 (b) Pass 2 (Affine) (c) Pass 2 (Spline)

(d) Pass 0 (e) Pass 2 (Affine) (f) Pass 2 (Spline)

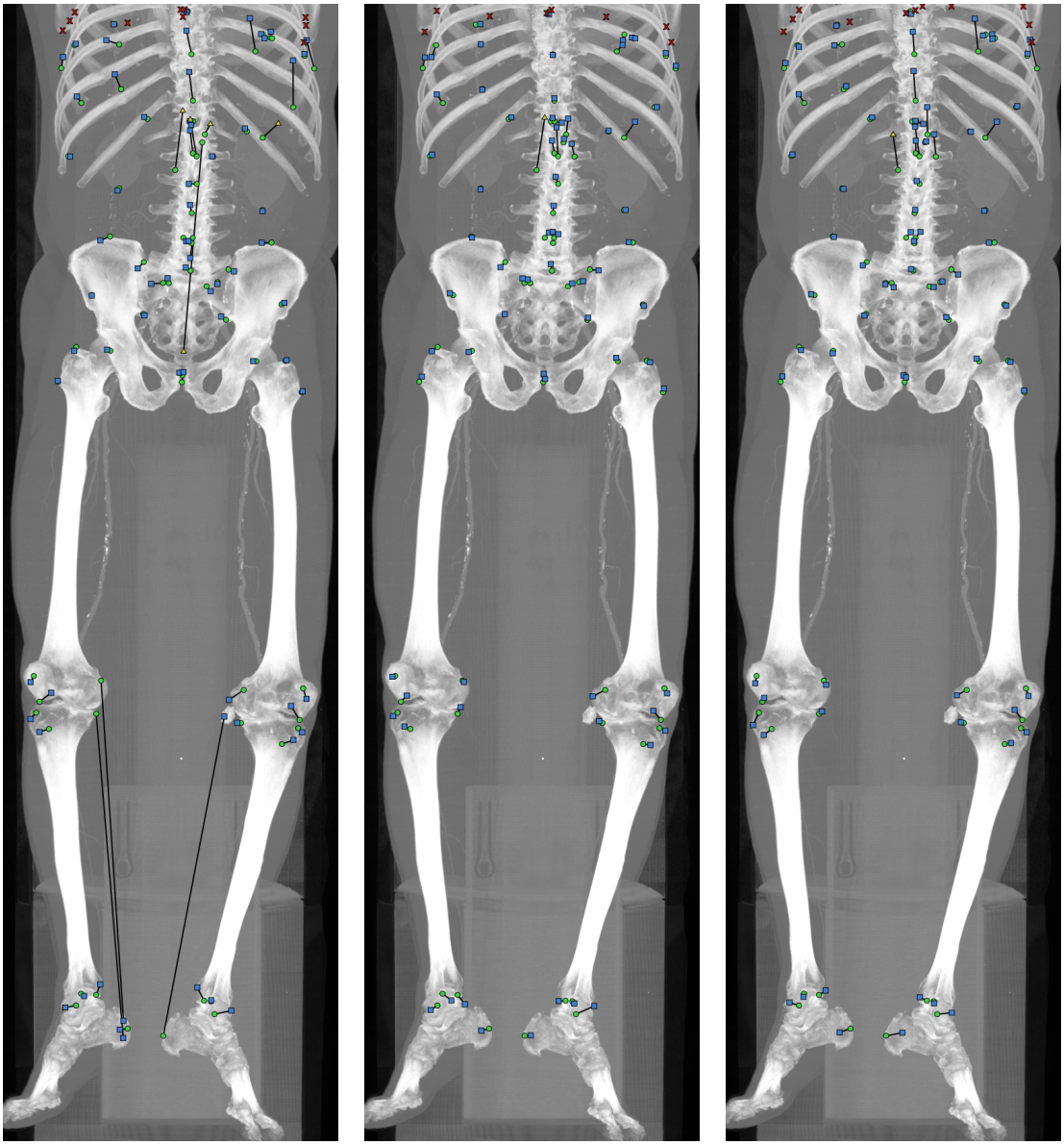
Figure 2.13: Coronal (top) and sagittal (bottom) MIP images of landmark detection results in a thoracic scan. The affine transform has corrected the ribs but the most inferior vertebrae are wrong. The spline has corrected about half of the rib and vertebrae landmarks. Other landmarks in the scan are generally already well detected in the zeroth pass. [TMVS Dataset ID: 3640]



(a) Pass 0 (b) Pass 2 (Affine) (c) Pass 2 (Spline)

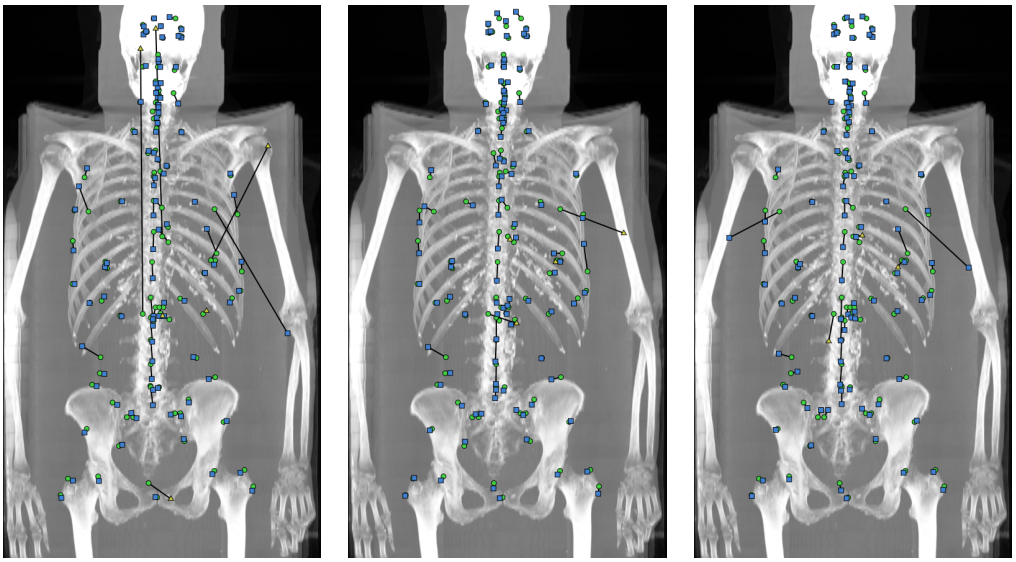
(d) Pass 0 (e) Pass 2 (Affine (f) Pass 2 (Spline)

Figure 2.14: Coronal (top) and sagittal (bottom) MIP images of landmark detection results in a cardiac scan. It can be seen that the vertebrae landmarks are corrected following the introduction of contextual information. Note that these landmarks are marked as uncertain in the ground truth, and any arrangement of detected vertebrae within one vertebra of the ground truth would be reasonable. The accuracy of the cardiac landmarks does not significantly change. [TMVS Dataset ID: 1433]

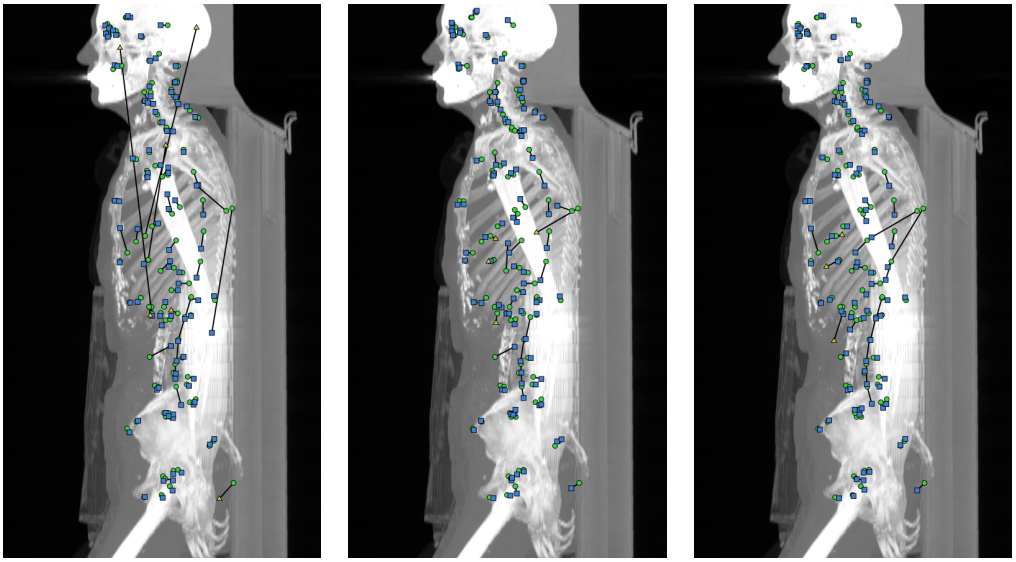


(a) Pass 0 (b) Pass 2 (Affine) (c) Pass 2 (Spline)

Figure 2.15: Coronal MIP images of landmark detection results in a scan of the lower limbs. This patient has turned out legs and feet. There is also an atypical bright mass on the medial side of the left knee. These variations have led to the knee and ankle joints being confused, and this is easily remedied using contextual information. The landmark at the origin of the superior mesenteric artery is also corrected from its initial position low down in the pelvis. Images © Vital Images, Inc., used with permission. [TMVS Dataset ID: 3197]

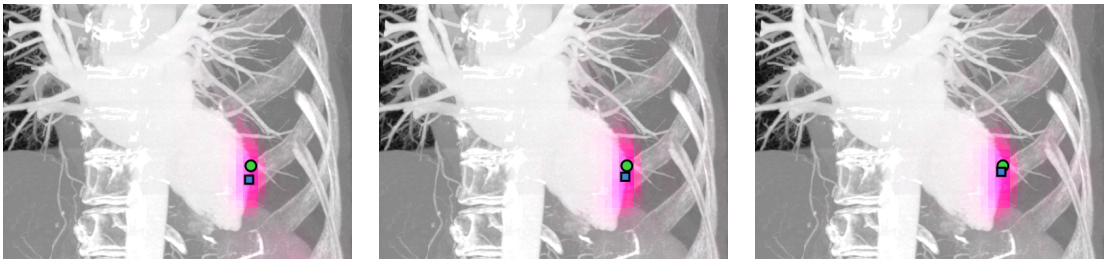


(a) Pass 0 (b) Pass 2 (Affine) (c) Pass 2 (Spline)



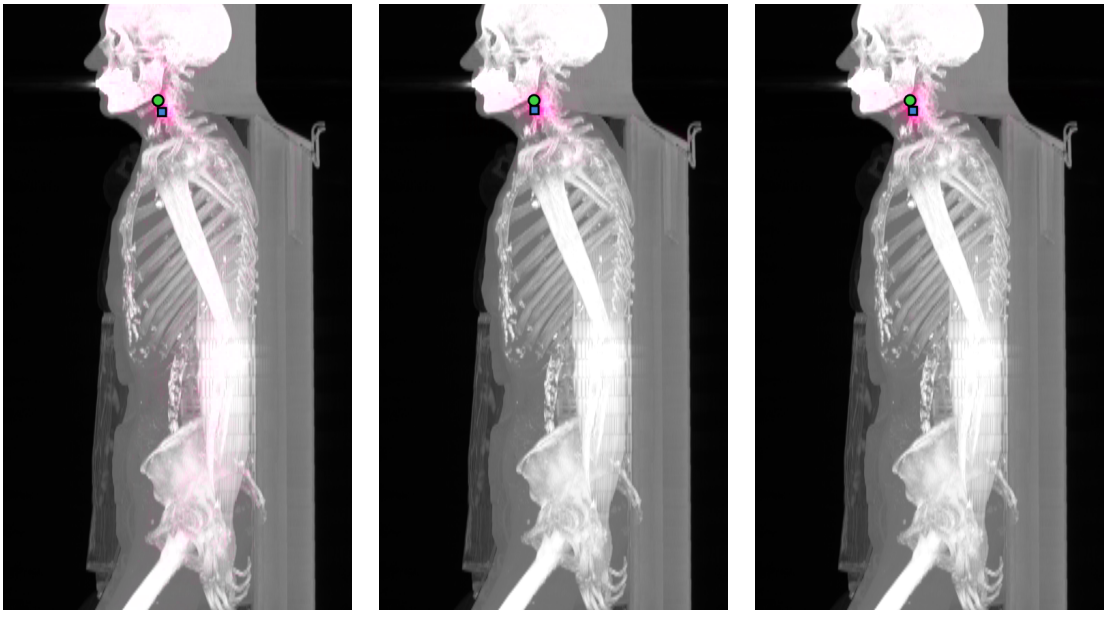
(d) Pass 0 (e) Pass 2 (Affine) (f) Pass 2 (Spline)

Figure 2.16: Coronal (top) and sagittal (bottom) MIP images of landmark detection results in a thoracic/abdominal scan. This dataset shows the correction, by atlas location autocontext, of cardiac and abdominal landmarks which have been detected in the upper thorax and head. The scapulae landmarks have been detected on the upper arms, and atlas location autocontext does not correct this. This will be because almost all datasets have the patient positioned with their arms above the head, which means that the scapulae are rotated outwards and the arms are not visible in the scan. Many of the vertebrae have been detected on the adjacent inferior vertebra. The likely cause is that in this patient, there is an extra lumbosacral vertebra (traditionally vertebrae are marked from the top down and the lowest vertebra is called L6) and so the vertebrae, as marked in the ground truth, are located higher relative to other structures than they would be in standard anatomy. [TMVS Dataset ID: 1501]



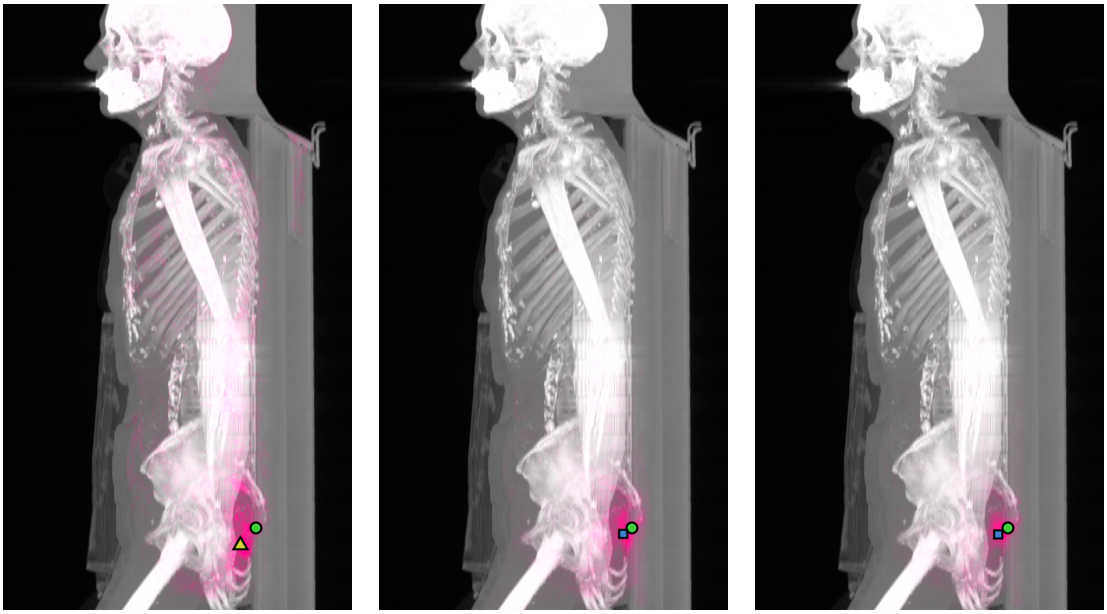
(a) Pass 0 (b) Pass 2 (Affine) (c) Pass 2 (Spline)

Figure 2.17: Probability cloud for an example of landmark: *Heart apex (extrema in sagittal plane) at endocardium*. [TMVS Dataset ID: 1433]



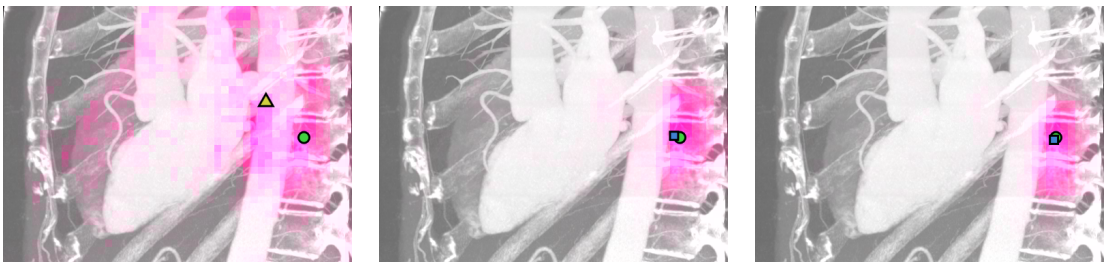
(a) Pass 0 (b) Pass 2 (Affine) (c) Pass 2 (Spline)

Figure 2.18: Probability cloud for an example of landmark: *Bifurcation of left common carotid artery into left internal and right external carotid arteries*. [TMVS Dataset ID: 1501]



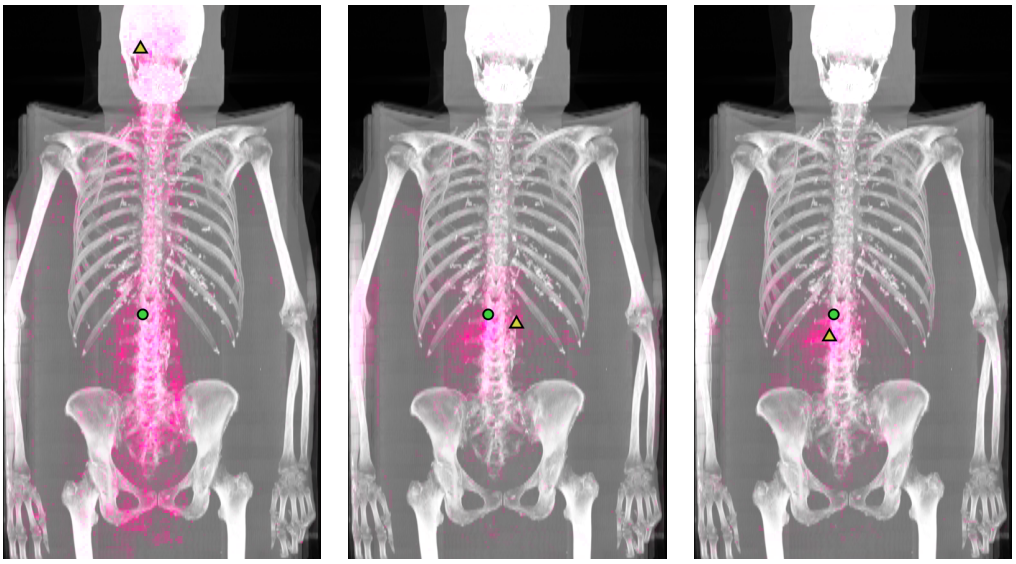
(a) Pass 0 (b) Pass 2 (Affine) (c) Pass 2 (Spline)

Figure 2.19: Probability cloud for an example of landmark: *Tip of the coccyx*. [TMVS Dataset ID: 1501]



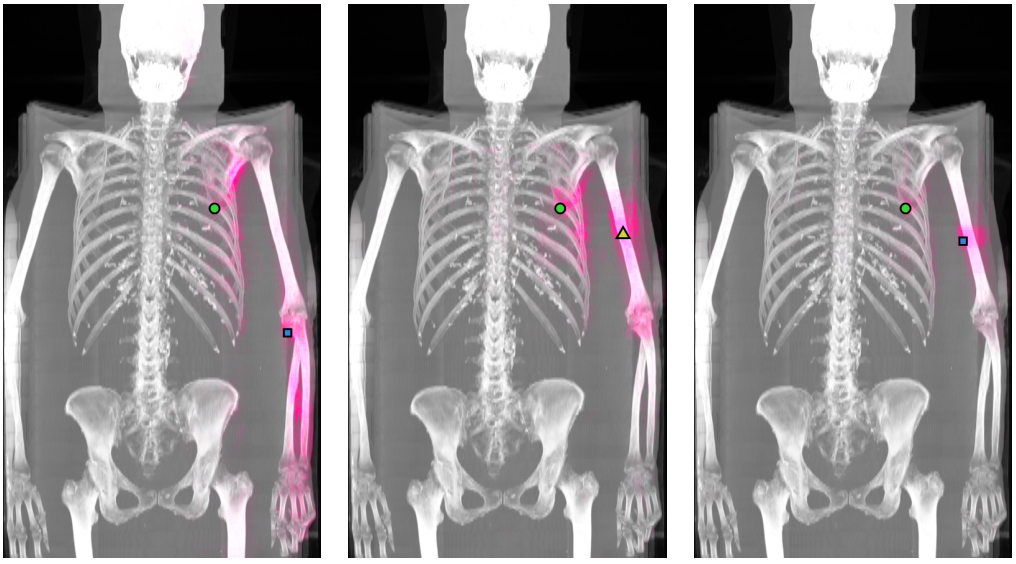
(a) Pass 0 (b) Pass 2 (Affine) (c) Pass 2 (Spline)

Figure 2.20: Probability cloud for an example of landmark: *Centre of body of T9*. This is an "uncertain" landmark. It can be seen that this cloud is more diffuse than those of the previous figures, extending over two or three vertebrae. As with all rib and vertebrae landmarks, there is some ambiguity between adjacent instances, which it appears the atlas location features do not fully resolve. [TMVS Dataset ID: 1433]



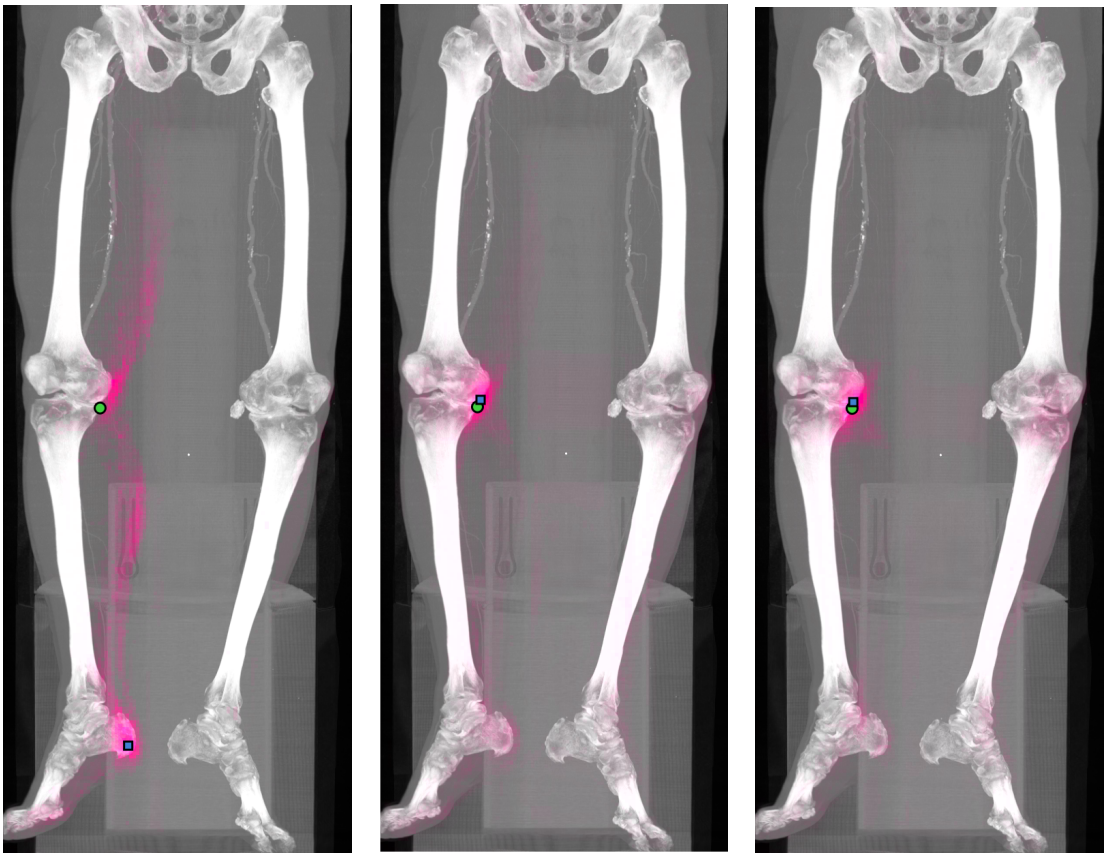
(a) Pass 0 (b) Pass 2 (Affine) (c) Pass 2 (Spline)

Figure 2.21: Probability cloud for an example of landmark: *Head of pancreas*. This is an "uncertain" landmark. [TMVS Dataset ID: 1501]



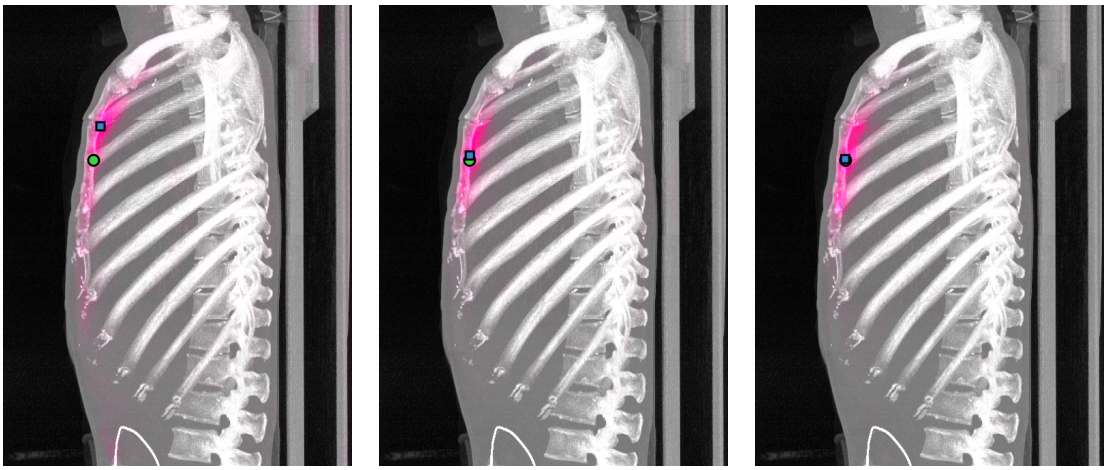
(a) Pass 0 (b) Pass 2 (Affine) (c) Pass 2 (Spline)

Figure 2.22: Probability cloud for an example of landmark: *Inferior angle of left scapula*. There is the unwanted side effect in pass 1 (both mappings) that a clear probability cloud materialises on the upper arm, for the reason, as mentioned in Figure 2.16, that almost all datasets have the patient positioned with their arms above the head, with the scapulae rotated outwards and the arms not visible. More datasets with this posture are required as training examples. [TMVS Dataset ID: 1501]



(a) Pass 0 (b) Pass 2 (Affine) (c) Pass 2 (Spline)

Figure 2.23: Probability cloud for an example of landmark: *Medial condyle of right tibia*. [TMVS Dataset ID: 3197]



(a) Pass 0 (b) Pass 2 (Affine) (c) Pass 2 (Spline)

Figure 2.24: Probability cloud for an example of landmark: *Costal cartilage junction of 3rd rib left side*. [TMVS Dataset ID: 3640]

2.5 Discussion

2.5.1 Summary of our contribution

We propose atlas location autocontext, a lightweight method of exploiting spatial information for standard feature-based machine learning algorithms. Atlas location autocontext is a good fit in the context of landmark detection, and brings additional benefits of robustness, fast computation time and low memory requirements. We have demonstrated both qualitatively and quantitatively that use of this mechanism improves both the accuracy and the detection reliability of an anatomical landmark detection algorithm.

The mean landmark error that we achieve, of slightly less than 10mm, is still too large for many clinical applications, although some landmarks are detected better than others, as illustrated in section 2.4.3. There are a number of sources of residual error which we discuss below, and which we will go on to address in later chapters.

2.5.2 Mapping accuracy versus autocontext utility

The mapping accuracy evaluation in section 2.3.3.3 showed the thin plate spline to be significantly more accurate than the affine transformation. However, visualisation of the landmark probability clouds shows that similar spatial information is being extracted, and the numerical results support this. If anything, the affine transformation performs better for the purpose of atlas location autocontext, giving greater improvement in AUC.

To investigate this, we ran a landmark detection experiment using the *ground truth* landmarks for the mapping (only one iteration required), in addition to the standard intensity features. In theory, when the spline mapping is used, we are providing perfect information about the true location of the landmarks, and the forest should be able to learn and return perfect results. In practice, the results are as shown in Table 2.5.

As expected, the spline gives much better performance than the affine mapping. The results for the spline are not perfect — and in fact worsen when intensity features are included — which we can put down to:

- The lower resolution at which detection is run.
- The class labelling strategy in which spheres rather than points are used to represent the landmarks.

Mapping	Features	Mean error (mm)	AUC
Affine	$T_a(v)$	12.71	0.890
Affine	$T_a(v)$, $d_{sag}(v)$	12.66	0.897
Affine	$T_a(v)$, $d_{sag}(v)$, $I(v+d)$	8.17	0.955
Spline	$T_a(v)$	2.36	0.976
Spline	$T_a(v)$, $d_{sag}(v)$	2.43	0.980
Spline	$T_a(v)$, $d_{sag}(v)$, $I(v+d)$	3.64	0.990

Table 2.5: Results of landmark detection using mappings created from the ground truth. Spline = Thin plate spline (similarity). Results are the means of three experiment runs, run with different randomisation seeds.

- The training sample selection strategy where background samples are picked uniformly across the volume. It might be necessary to sample more densely close to the landmarks in order to force the trees to discriminate between landmarks and voxels close in location.
- The training sample selection strategy where the majority of the training sample population is background voxels (which are each given as much or greater weight than the landmark samples). Spatial splits provide little information with respect to background samples, so in the first place intensity features are likely to be chosen in preference to atlas location features.

Excepting the point about low resolution, these issues are artefacts of the fact that landmark detection has been framed as a *classification* problem. We note here that a different model, perhaps a regression forest such as that of Gao and Shen [18, 55], might be a more natural fit. In the limit where the atlas features are derived from ground truth, the forest would simply have to learn the identity regression function. In contrast, in the classification formulation, there is still a reasonably complex function to learn. This is considered in future work.

This experiment gives us confidence that the code is correct, and that provision of accurate information about location leads to better localisation of landmarks by the forest - as it should! However, when detected landmarks are used for the atlas coordinate features, the more accurate thin plate spline no longer leads to better localisation, and in fact the affine mapping performs slightly better. We explain this by the fact that the choice of measurement metric for mapping accuracy was not considered in the light of the autocontext role which this

mapping would play. We made no distinction between landmarks which were already well detected by the forest, and landmarks where there was room for improvement. It is obvious that the spline mapping preserves accuracy much better in the former case; however, this will not translate into an improvement in detection accuracy. Poorly detected landmarks which are *positively* detected will be mapped *exactly* to their counterparts by the spline, unless they fall below the iterative fitting error threshold τ_E. On the other hand, the affine transformation is likely to predict a more reasonable location. Landmarks which are *not* detected will be better interpolated by the spline than the affine transformation, *if* the surrounding positively detected landmarks are themselves well detected. If the surrounding landmarks are poorly detected, and the errors are not too strongly correlated, then the affine transformation may actually provide a better predicted position since errors will be somewhat cancelled out. See section 2.6.1 for further discussion of this.

2.5.3 Examination of the remaining sources of error

2.5.3.1 Inter-subject variation

As with all image analysis algorithms, inter-subject variation is a problem. Anatomy may be atypical or not visible in the scan, due to normal anatomical variation or pathological variation, leading the identification and even the definition of a landmark to fall down.

Both ground truth annotation and detection for atypical anatomy is problematic. If a specific variant is of interest, then a ground truth strategy must be adopted and a reasonable number of training examples are required for this flavour of landmark appearance to be learnt.

2.5.3.2 Rib and vertebrae repeating structures

Another major category of errors relates to the repeated rib and vertebrae structures. All too often, landmarks are detected on neighbouring structures. Firstly, there is little to differentiate these structures in terms of appearance. Secondly, the position of neighbouring soft tissue organs varies between subjects and between phases of respiration. In some instances, this is an insoluble problem, such as in the cardiac dataset of Figure 2.14, where the scan does not contain any uniquely identifiable vertebrae, so even the ground truth cannot be reliably created. However, other cases are more tractable.

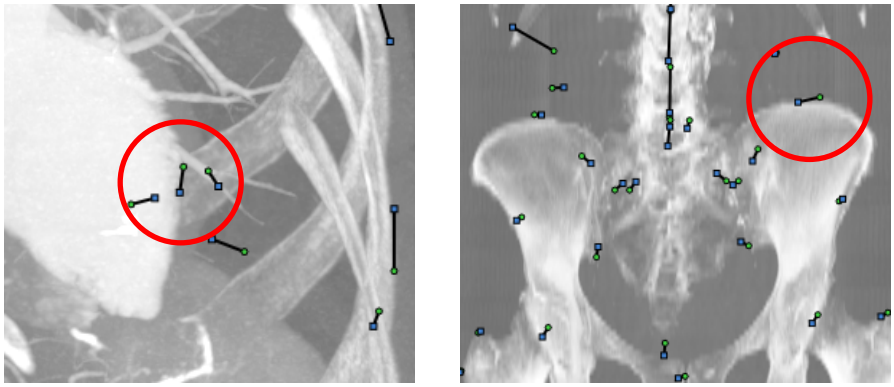
The constraints introduced by atlas location autocontext are soft constraints, too soft for the scale at which the repeating structure problem manifests. In addition, errors are often correlated i.e. many ribs lie "one up" from the true position. Where the number of false positives outnumbers the number of the true positive results, the mapping will be driven by the false positive results, at least according to the current mapping generation method.

It may be that the reliability of rib and vertebrae landmarks (and indeed other landmarks) could be learnt in order to inform the choice of detected landmarks which contribute to atlas registration. Alternatively a specific post-processing step may be necessary to constrain these repeating structure landmarks to anatomically plausible configurations. For instance, a statistical shape model [13], or a graphical model as in [16, 61]. There have been a series of recent MICCAI workshops looking specifically at computational methods for spine imaging [62, 63, 64].

2.5.3.3 Ill-defined and surface landmarks

There remains a general problem with precision. Visual comparison of the detected landmarks with the ground truth suggests some landmarks have been poorly defined. This is particularly the case for surface landmarks. These are landmarks on a gentle curve or gently curved surface. Their local appearance is close to a line or a plane, meaning that there are many similar looking voxels in the vicinity of the landmark, which poses a problem for exact localisation. Two examples are shown in Figure 2.25. The detected positions seen here are not unreasonable, and perhaps there is a range of acceptable "error" in these cases. Indeed, at TMVS we have done work (to which I have contributed) on the evaluation of registration by measuring errors only in the direction of the surface normal [65], on the basis that manual landmarking is possible with greater accuracy in this direction than in the surface plane.

Further, for ease of marking, many surface landmarks are defined with reference to the planes of the volume (e.g. most superior or lateral point, using the volume-aligned axial or sagittal plane), which is itself aligned with the scanner table, and not with reference to anatomical structures. This makes landmark positioning sensitive to changes in patient orientation or posture. In Figure 2.25b which shows the left iliac spinal landmark, it can be seen that the tilt of this patient's pelvis has led to a shift in the landmark, and the detected landmark is closer to the no-tilt ground truth position.



(a) (b)

Figure 2.25: Figure showing close-ups of the affine atlas feedback results for two landmarks a) Sagittal MIP image of *Heart apex [extrema in sagittal plane] at endocardium* and b) Coronal MIP image of *Superior aspect of left iliac spine*. These are surface landmarks where the landmark is poorly defined in the plane of the surface. Both landmarks are defined with reference to the orientation of the patient in the scanner, and not with reference to anatomical structures. In the case of the iliac spine, the pelvis is tilted to the left which has shifted the ground truth landmark to the right. However, the detected landmark is close to the no-tilt ground truth position. [TMVS Dataset ID: 1501]

As a result of this analysis, we have since done work at TMVS on developing a redefined and expanded set of landmark ground truth.

- Ill-defined landmarks have been removed.
- Landmarks have all been defined with reference to anatomical structures rather than volume planes.
- Landmarks have been defined on the whole body (arms, hands and feet were previously excluded).
- Gender-specific landmarks have been introduced.
- There is a greater density of landmarks e.g. previously we were only locating the odd-numbered thoracic ribs and vertebrae.

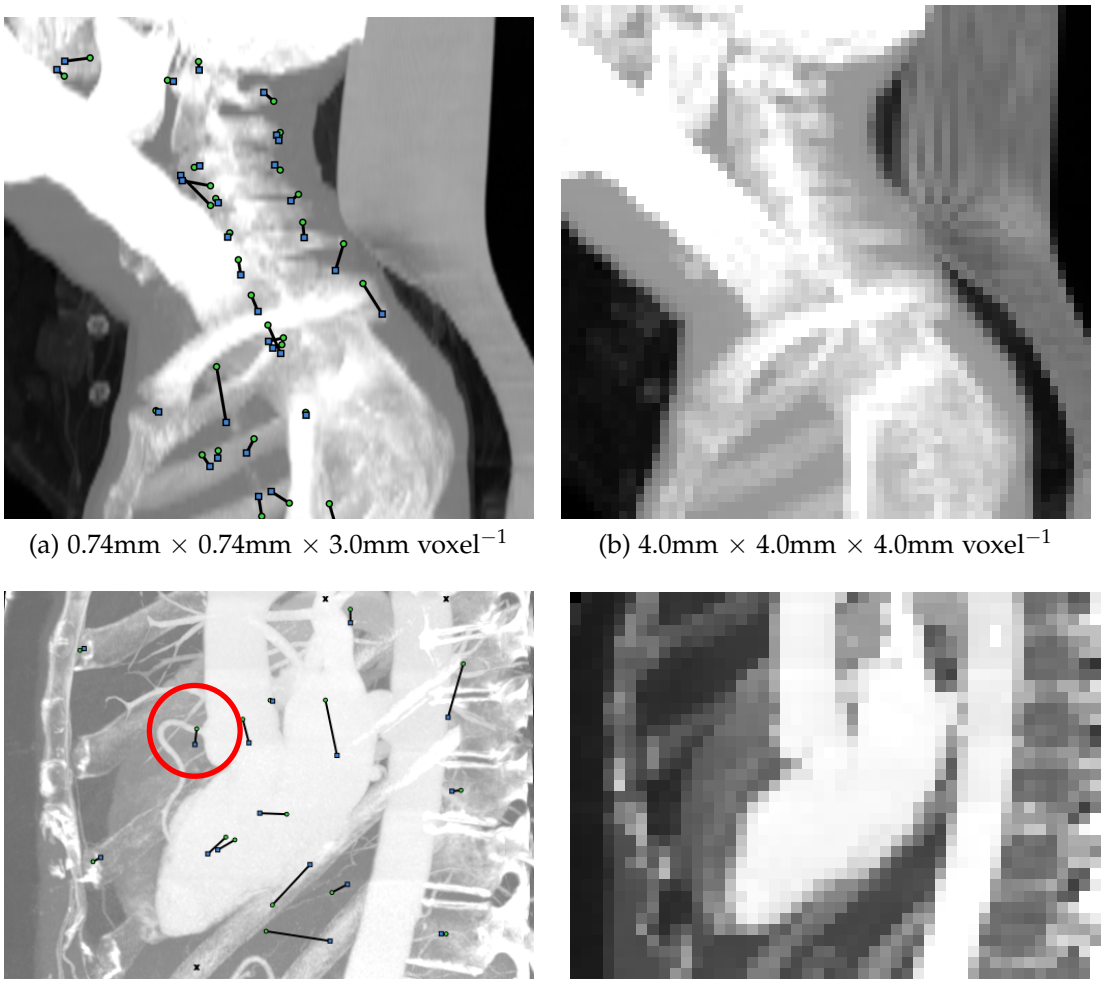
We use the head and neck subset of these landmarks in the experiments in chapter 3.

2.5.3.4 Low operating resolution

A second aspect which cannot be ignored when discussing precision shortfalls is the fact that we are running the algorithm at a low data resolution of 4mm

voxel^{-1}. This low resolution is being used in order to maintain a speed close to real time. Two examples are shown in Figure 2.26.

In section 5.3 we investigate the use of higher resolution for vascular landmarks, showing this does indeed give a significant improvement in accuracy.



(a) 0.74mm × 0.74mm × 3.0mm voxel^{-1}

(b) 4.0mm × 4.0mm × 4.0mm voxel^{-1}

(c) 0.37mm × 0.37mm × 0.4mm voxel^{-1}

(d) 4.0mm × 4.0mm × 4.0mm voxel^{-1}

Figure 2.26: Figure showing the problem that the low resolution is causing for localisation of landmarks which are on very fine structures. On the left are sagittal MIP images of the original volumes, overlaid with the pass 2 affine transform results. On the right are sagittal MIPs of the volumes used for training and detection. a) and b) show the neck. In the downscaled image the separation between vertebrae is much less distinct, as are the tips of the spinous processes of the cervical vertebrae. c) and d) are from the cardiac dataset shown in Figure 2.14, and the red circle highlights the right coronary ostium (the point on the ascending aorta from which the right coronary artery originates). In the downscaled image, the right coronary artery cannot be made out because the diameter of the coronary arteries is less than 4mm. [TMVS Dataset IDs: 1627, 1433]

2.6 Future work

2.6.1 Improving the atlas mapping

2.6.1.1 Landmark reliability and weighted least squares fitting

In the computation of the affine transformation, we have taken no account of the variability in accuracy and reliability of the different landmarks, beyond a simple binary thresholding according to the forest probabilities (at $P_F \geq \tau_P$) to separate true positives from false positives. We could, however, have chosen to relate the landmark probabilities directly to their mapping contribution, using *weighted* least squares fitting. We add the weights w_i, $i = 1 \ldots N$ to equation 2.18.

$$T_a(v) = \underset{T_a(v)}{\arg\min} \sum_{i=1}^{N} w_i |l_i - T_a(l_i)|^2 \tag{2.24}$$

As suggested above, we could simply assign the weights to be the raw probability values $w_i = P_F(l_i|c_i)$. However, weighted least squares fitting is a known solution to the problem of heteroskedastic data. If we knew the variance of each landmark error, we could assign the reciprocal to be the weight, $w_i = \frac{1}{\sigma_i^2}$, and the overall solution would be the best linear unbiased estimator [66]. There is a caveat that this assumes *uncorrelated* errors —- this would have to be a pragmatic assumption, since we know it is not always the case. In fact, we can estimate the variances easily from the training data results (see Figure 2.12). For the purpose of the mapping, we would want to measure the variance of only those landmarks with $P_F \geq \tau_P$. Given enough data, we could fit a linear (or more complex) model relating error variance to forest probability, since we might expect that landmarks with a higher probability have lower variance.

2.6.1.2 Registration by parts: Bringing anatomical knowledge into the equation

Thus far, we have used generic mapping techniques, taking no account of anatomical knowledge. We highlighted that the affine transform has the nice property of error-cancelling for regions where the landmarks have random uncorrelated errors. However, basic knowledge of skeletal articulation tells us that a global transform may fit a single body part in a standalone scan better than it would that body part in a larger scan containing multiple articulating body regions, which are known to move quasi-independently, and vary in size or length quasi-

independently e.g. head size or leg length.

It may be worth investigating a locally affine approach which treats the body as a few distinct structures. One possible division of the current landmarks is: head, neck, thorax & abdomen, legs. These regions could be mapped locally using an affine transform, and then connected into a single mapping using a spline or by other means. This would be done by taking the estimated atlas locations of the landmarks, according to the partial affine mappings — which would not lie exactly at the atlas landmark locations — and mapping the detected landmarks to these using a single spline. We expect that using a locally affine approach would give better mapping accuracy in the head, and enable atlas location feedback to produce an improvement in the head and neck landmark errors.

Alternatively, we could take a statistical shape modelling approach [13] and identify modes of variation using a method such as principal component analysis (PCA). In this way, we might *implicitly* discover spatial behaviour such as articulation and the differing movement of bones and soft tissues. To prevent the mapping becoming too flexible, the number of modes of variation and the size of the valid shape domain would require to be limited.

2.6.2 Expanding the atlas coordinate features

The theme of using atlas registration to generate new image features could be expanded. In Table 2.6 we provide some examples of possible spin-off features. At the least, we could use the affine scale components to normalise the standard intensity feature offsets, to correct for differences in scale between smaller and larger patients in the study population — which may, in future research, include children.

2.6.3 Introducing prior information

We mentioned earlier that this algorithm is intended as a general purpose tool which could be applied to any and all scans exiting a scanner. Hence, no assumptions are made about the protocol or content of the scan, and we train on a diverse population of datasets.

However, if the clinician provides an indication of the body region and pose of the patient being scanned, then the atlas coordinate features could be used to train a zeroth iteration which has knowledge of an approximate atlas space mapping. For example, a simple translation operation could be used to centre

Feature	Description
$T_a(v)$	The original atlas coordinate of a point v.
$T_a(v+d)$	Atlas coordinate of a neighbouring voxel at a displacement d.
$\lvert T_a(v) - p_a \rvert$	Distance in atlas space from a specified atlas coordinate p_a i.e. describes a spherical boundary in atlas space.
$\frac{d}{dx}(T_a(v+d))$, $\frac{d}{dy}(T_a(v+d))$, $\frac{d}{dz}(T_a(v+d))$	First order derivatives (or gradients) in x, y and z of the spline warp field at $v+d$. These could be efficiently estimated by central difference.
$f(T'_a(T_a(v)+d))$	Standard feature value f (e.g. image intensity I) of a neighbouring voxel in atlas space. $T'_a(v)$ represents the mapping from atlas space to volume space.
$f(v + s_{Ta} * d)$	Standard feature value f (e.g. image intensity I) of a neighbouring voxel using an offset d which has been scaled by the scaling component s_{Ta} of the atlas space. We could go further and downscale the volume to a resolution specified in atlas space rather than in scanner space. The idea of exploiting scale is to compensate for differences in patient size (e.g. the discrepancy between a tall man and a small woman may be large).

Table 2.6: Features based on the atlas mapping concept, for future experimentation.

the volume in atlas space at the relevant body part. This would support the goal of maintaining a very fast detection time. It would also reduce the number of instances of complete failure cases, which can be a particular problem for scans with very limited acquisition regions and thus limited contextual information, for instance scan of a single hip or knee joint.

By the same token, if other patient information were made available, such as height, weight, age and gender, this could also be incorporated. Firstly, by providing constraints on the mapping component values — constraints for which the body region and pose would also, self-evidently, be highly informative. Secondly, explicit features could be introduced and made available to the forest, taking the same value for all voxels in a given scan e.g. $f_{height}()$, $f_{weight}()$, $f_{age}()$ and the binary-valued $f_{gender}()$.

Chapter 3

Anatomical landmark detection using gradient orientation features

Abstract

This chapter moves anatomical landmark detection into the realm of MRI data, focusing on head scans. As in chapter 2, target scans may come from any protocol and any scanner, and so we seek a general-purpose algorithm that is robust to differences in scan resolution, plane of acquisition, intensity value distributions, field of view and degree of noise present. Unlike CT data, MRI data is uncalibrated and so intensities vary widely scan to scan, the distribution of values varying both in range and in shape. We motivate the use of histograms of unweighted — and in the case of cross-modality detection, unsigned — gradient orientations, as an alternative set of features to simple intensity features. By discarding the spatial and magnitude weighting schemes that are used in the SIFT and HOG feature descriptors, we arrive at an efficiently computed feature which is robust to intensity invariances. To distinguish structural detail from noise, we rely on the aggregation of orientations over a spatial region. Extensive parameter exploration experiments are presented. We further characterise the trade-offs between run time speed, accuracy and memory requirements, which show somewhat surprising results since the forest shortcut mechanism described in 2.3.2.6 means that a forest of larger trees (i.e. more datasets per tree) may be faster at run time than a forest with the same number of smaller trees. Finally, we show how unsigned *gradient orientation features may enable cross-modality anatomical landmark detection, and how gradient orientation features slightly improve the accuracy of the whole-body CT detector of 2 when used alongside intensity features.*

3.1 Synopsis

In this chapter we

- (3.2.3) Motivate the use of histograms of oriented gradients for application to anatomical landmark detection in medical scans, specifically in MRI head scans.
- (3.4.2 - 3.4.6) Investigate aspects of gradient orientation features: the sampling strategy for the cuboid sizes and offsets over which the histograms are computed, the number of histogram bins and a Gaussian bin weighting scheme, the plane in which the 2D gradient orientations are computed and how this correlates with the plane of scan acquisition, and finally the use of noise thresholding.
- (3.4.7) Compare gradient orientation features with simple intensity features, showing that gradient orientation features work better in MRI volumes, and that intensity features work better in CT volumes.
- (3.5) Illustrate the trade-offs between accuracy, run time and memory usage. Show how these can be characterised (data sheet style) on a graph plot.
- (3.6) Evaluate the improvement that HOG features confer on the whole-body CT landmark detector from chapter 2.
- (3.7) Show that a detector may be trained on one modality and applied to another (cross-modality classification), using unsigned gradient orientations.

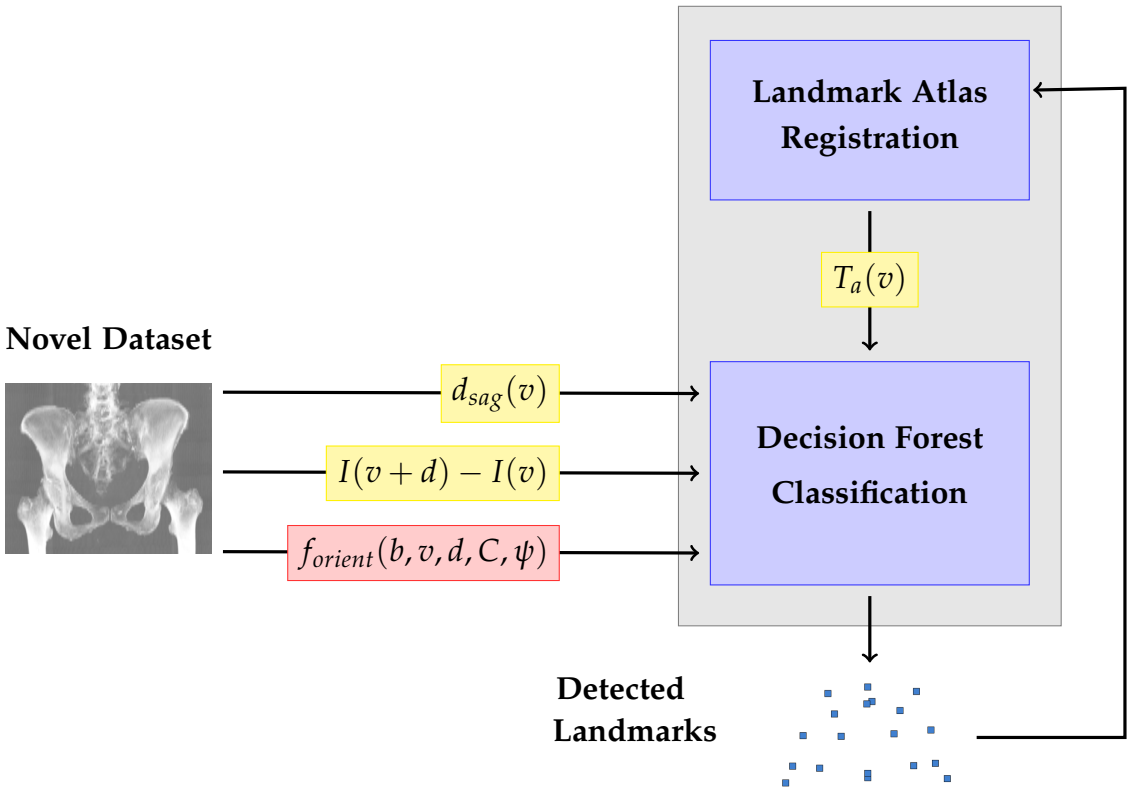


Figure 3.1: Simplified overview of anatomical landmark detection (see Figure 2.3 for full diagram). In this chapter we focus on the new set of image features $P(b|v+d, C)$ marked in pink. These features are employed during landmark detection in addition to the sagittal displacement feature $d_{sag}(v)$, the relative intensity features $I(v+d) - I(v)$ and the atlas location features $T_a(v)$. The new gradient orientation feature is the probability of a voxel gradient being oriented within the angle range of bin b, in the cuboid of size s lying at an offset d from the voxel of interest v.

3.2 Introduction

3.2.1 Problem definition

In chapter 2, low level image intensity features were used. It is anticipated that performance gains may be obtained through the use of higher level features. In particular, we look to identify a robust set of features for modalities other than CT, such as MR and ultrasound imaging, in which image intensity values are uncalibrated. The intensity distributions of two MRI-T1 series images may have completely different spreads and shapes; we cannot assume a linear relationship between the intensities.

3.2.2 Prior art

Authors have tackled the problem of analysing uncalibrated data in different ways. Some different approaches are outlined below.

A model of the imaging process may be used to inform interpretation of intensities. In 2004 Fischl *et al.* [67] modelled the physics of MRI and came up with a method to estimate the intrinsic tissue parameters (T1, T2* and proton density) from a scan, for the problem of brain segmentation. This approach was developed for fast low-angle shot (FLASH) and spoiled gradient recalled echo (SPGR) sequences, for which the physics equations are straightforward, and required multiple scans to be acquired in order to estimate the (multiple) tissue parameters. Then, taking the estimated tissue parameters and *given* the MR acquisition parameters for the novel image to be segmented, the mean and variance for the difference tissue classes could be predicted. More recently, for the purpose of MR brain image synthesis, Jog *et al.* [68] proposed a related method which assumed typical intrinsic tissue parameters from the literature, used a fuzzy c-means algorithm to identify mean tissue intensities, and then solved for the acquisition parameters.

However, we often have little information as to the nature of the imaging technique or its acquisition parameters, and may not be able to make assumptions about the content of the scan (and therefore the dominant tissue types). Hence, we turn to feature descriptors which claim a degree of grey-scale invariance. One approach is to use ranking of image intensities. Intensity rankings are robust to monotonic transformation of the intensities. Local binary patterns (LBPs) were proposed by Ojala *et al.* [69] in 1996 and later extended to make a rotationally invariant operator for the purpose of texture classification [70]. In essence, LBPs are binary sequences of digits referring to the pixels in a local neighbourhood, thresholded at the value of the central pixel i.e. a '1' denotes that a pixel's intensity value is greater than that of the centre, and a '0' denotes that it is less. Liu *et al.* [71] later used principal component analysis to obtain dimension-reduced LBPs for the diagnosis of macular pathologies in optical CT images. Other authors have applied LBPs to face recognition [72], to texture classification in lung CT images [73] and to false positive reduction in mammographic mass detection [74]. Alternatively, the *ranklet transform* [75] is a multiresolution, multi-orientation image processing technique analogous to the wavelet, but which works with grey-scale rankings rather than absolute intensities. It has been applied to the problem of facial detection [75] and mammographic tumour detection [76, 77]. Yang *et al.*

[77] applied the ranklet transform prior to using grey-scale co-occurrence matrix (GLCM) features [78] for the purpose of texture measurement in mammographic images. GLCM features measure properties such as homogeneity and presence of linear dependencies. Images from three different scanners were used and it was found that the ranklet-transformed result outperformed the corresponding wavelet-transformed result because the former was more robust to acquisition differences.

An alternative approach is to move into the domain of gradient *orientations*, by which we mean the direction of the dominant gradient at any given pixel (or voxel). Freeman and Roth [79] developed the idea of using orientation histograms (based on a patent by McConnell [80]) and demonstrated these for hand gesture recognition in images [79, 81]. The idea is that a region can be characterised by counting the frequency of occurrence of each gradient orientation over all pixels in the region, to form a histogram. In this original version, gradients were thresholded at some chosen magnitude, below which measurements were assumed to be inaccurate, leaving only the edge gradients. Popular variants include the Histogram of Oriented Gradients (HOG) descriptor of Dalal and Triggs [82, 83, 84, 85, 86], and the Scale-Invariant Feature Transform (SIFT) of Lowe [87, 88, 89]. Both were proposed for the purpose of object recognition in 2D images; we describe some of the properties of these descriptors in section 3.2.3. Allaire *et al.* [90] have since extended the SIFT descriptor to three dimensions for use in medical imaging. Gradient information has also demonstrated value for the problem of medical multi-modality image registration [91, 92], where normalised gradient fields are used to compute a similarity metric which indicates how well corresponding image gradients are aligned.

Finally, we mention the *modality-independent neighbourhood descriptor* (MIND) proposed by Heinrich *et al.* [93] idea which measures the self-similarity (by means of sum of squared differences, SSD) of an image patch to a nearby patch at a defined offset, for any number of desired offsets and patch sizes. By comparing an image only to itself, this descriptor is robust to inter-modality differences in intensity distributions. By considering only the local neighbourhood, it will be robust to changes within the image itself. A follow-up descriptor which has greater robustness to noise was proposed [94], self-similarity context (SSC), in which self-similarity is measured between neighbourhood pairs of patches, as opposed to always comparing to the central patch of interest. Li *et al.* [95] introduced a related descriptor ALOST (autocorrelation of local structure) in which the autocorrelation of *structural* information is measured rather than

the autocorrelation of *intensity* information. The structure is represented by measuring the mean and standard deviation (congruency) of the signal phase. There remains the problem that different modalities may not show exactly the same anatomical detail (some anatomy may be visible in one and not another) and so different images cannot correspond. To solve this for the problem of registration, Ou *et al.* [96] suggested the concept of mutual saliency meaning that some voxels are weighted more heavily than others according to a map which evolves during the registration process. This could be an idea worth exploration.

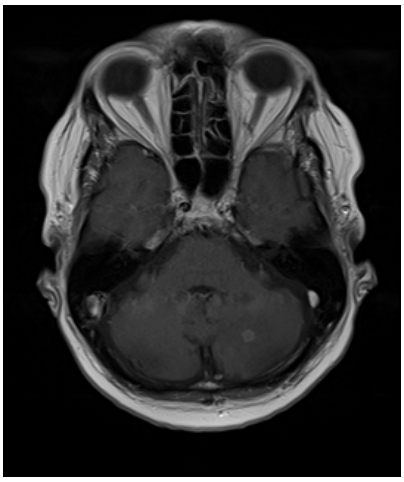
3.2.3 Motivation for our approach

We choose to investigate the use of features based on histograms of gradient orientation in the context of anatomical landmark detection. These are established features which can be efficiently implemented in 3D volumes. Efficiency is important in terms of speed and memory usage, both during training and testing.

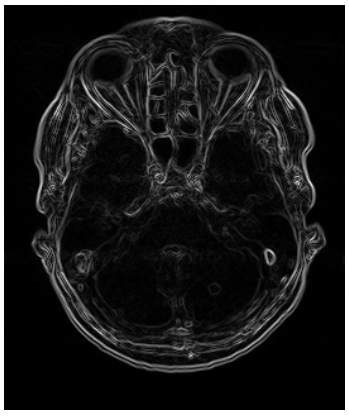
The utility of gradient orientation information may be intuited from Figure 3.2, which shows an example MRI-T1 slice from a head scan. The corresponding gradient magnitude and orientation information is shown, the latter with increasing degrees of quantisation. Even a binary representation is informative (see pictures 3.2f and 3.2g). Figure 3.3 shows example image patches and the corresponding histograms. It can be seen that the noisy image patch has a uniform histogram representation, whereas the edge patch has two histogram peaks at the two (opposing) edge directions.

In applying the HOG descriptor to people detection in 2D images and videos [82], unsigned gradients (i.e. 180-degree range) gave better performance than signed gradients (i.e. 360-degree range), due to the wide range of human clothing and background colours. Colours are fairly consistent in medical scans from the same modality. However, for the purpose of comparing scans of different modalities, where tissue intensities may be inverted, we may choose to discard the direction of the orientation and work with unsigned gradients. This results in a less powerful but more general descriptor. We investigate the use of unsigned gradients for cross-modality landmark detection in section 3.7.

A solution is desired for volumes rather than images. The obvious route when moving from 2D images to 3D volumes is to move from 2D to 3D gradients. This was done by by Allaire et al [90] who used 2D binning of gradients, by the azimuth and elevation angles, using 8 and 4 bins respectively of $\pi/4$ width. This method of division actually leads to two different sizes of bin. Equal sizes

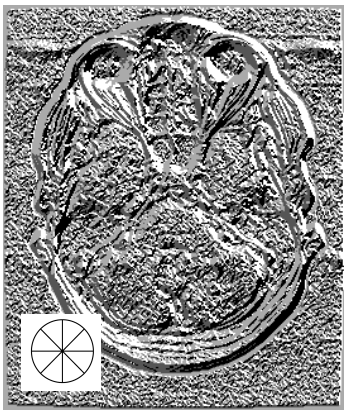


(a) MRI-T1 Axial Slice



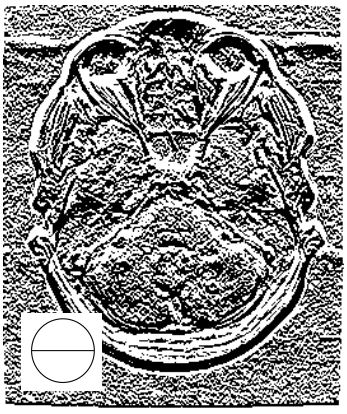
(b) Gradient Magnitude

(c) Gradient Orientation

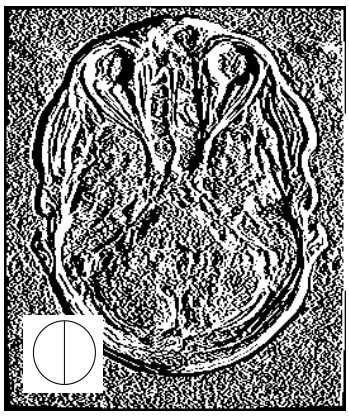


(d) 8 bins

(e) 4 bins



(f) 2 bins, horizontal intersect



(g) 2 bins, vertical intersect

Figure 3.2: Example head MRI-T1 axial slice, with the corresponding gradient magnitude and orientation images. Images d) to g) show the gradient orientation image when quantised into different numbers of bins. The orientation values are informative, even when severely quantised into two bins. Notice that there is amplification of the noise and of the horizontal imaging artefacts (which are of too small magnitude to be visible in the original image). [TMVS Dataset ID: 4660]

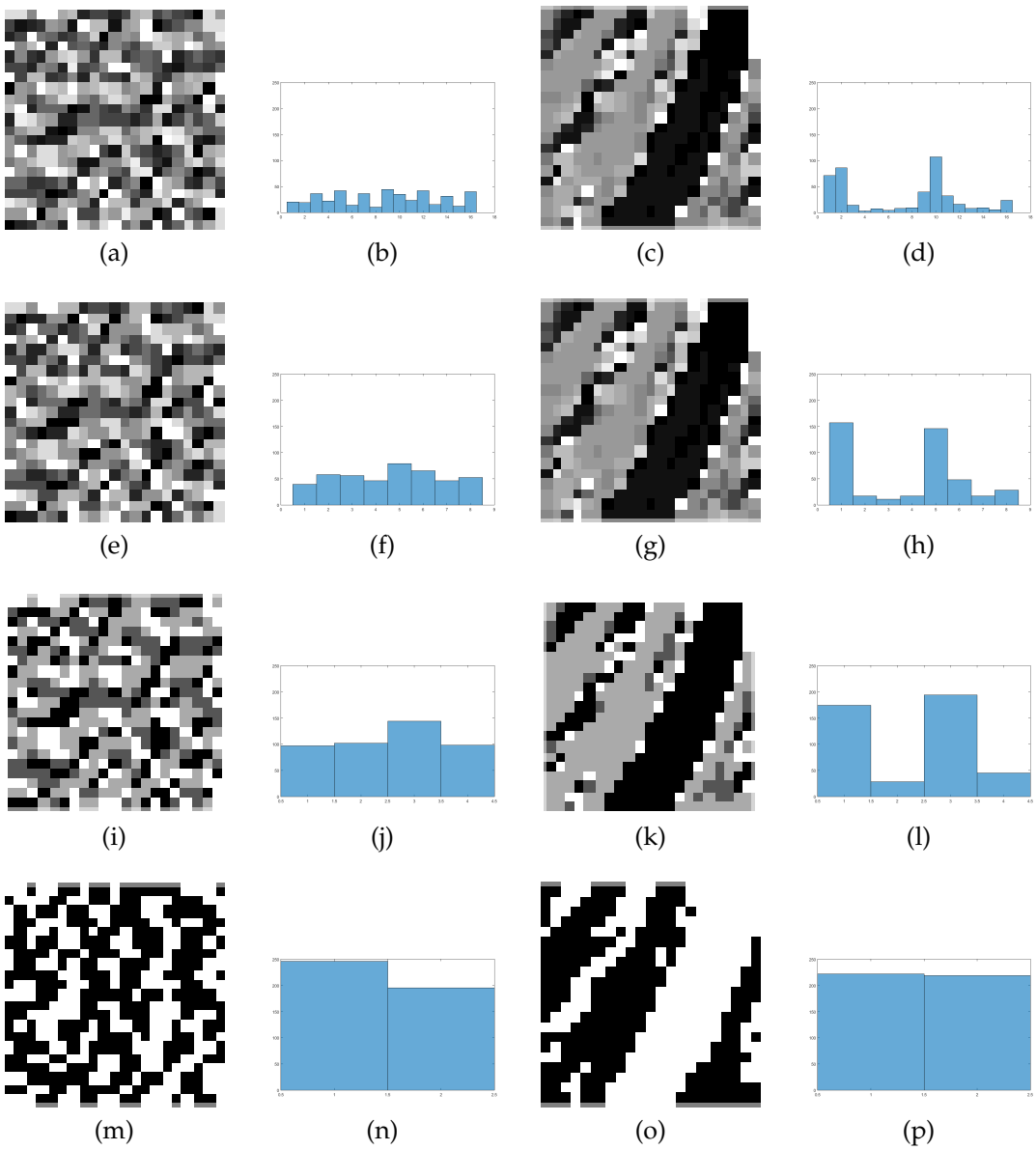


Figure 3.3: Images illustrating histograms of oriented gradients for different numbers of bins. From top to bottom row, the number of bins is {16,8,4,2}. We contrast a region of noise (left) with a region containing a strong edge feature (right). The patches are chosen from the head MRI-T1 slice of Figure 3.2. All histograms have the same scale on the vertical axis. [TMVS Dataset ID: 4660]

of bin could be achieved using alternative partition methods such as regular polyhedrons as was done by Kläser *et al.* [86]: the tetrahedron (4-sided), the cube (6-sided), the octahedron (8-sided), dodecahedron (12-sided) and the iscosahedron (20-sided). For illustration, imagine a 6-binned histogram created by centering a cube at the position of the gradient: the cube face out of which a ray directed along the orientation of a gradient emerged would be the relevant orientation bin.

In fact, we choose to work with 2.5D features, meaning that two-dimensional gradients are measured separately in the three anatomical planes (sagittal, axial and coronal), however a 3D feature offset is used and the histogram is computed over a cuboidal region. The theory is that sometimes MRI volumes have quite large slice spacing (compare the range of slice spacings in the CT data cohort in Figure 2.6 to that of the MRI data cohort in 3.7), and the gradients which are measured in the plane of scan acquisition will be more reliable than those in the normal planes (see investigation in section 3.4.5). By using 2D gradients we retain precision in the plane of acquisition. This way we can mix volumes with variable slice spacing which are acquired in the same plane.

There is unlikely to be a one-to-one correspondence between the orientation representations for different modalities, since often one imaging modality captures detail which is not visible on another. Perfect correspondence between images is also thwarted by the presence of noise. Noise is typically of very small magnitude, but in homogeneous regions such as air, it may be the determining orientation factor. Since we ignore magnitude information, there is no way to distinguish "true" gradients from noise. If noise has a truly random distribution (which should be the case for MRI images in which the noise distribution has been found to be approximately Gaussian where the signal-to-noise ratio is greater than two [97]), then we obtain a flat histogram which can be distinguished from spiky histogrammed regions of interest. However, minor artefacts may confound (see the horizontal artefacts in Figure 3.2). In section 3.4.6, we experiment with automatic detection of a noise threshold such as was used by Freeman and Roth [79]. This is analogous to the parameter ϵ employed in normalised gradient fields to prioritise real edges over noisy edges. At the other end of the spectrum, in their extension of SIFT to 3D for the purpose of medical image analysis, Allaire *et al.* [90] suggest introducing an upper threshold when working with CT images in order to remove background objects.

In the HOG and SIFT descriptors, gradient orientations are weighted by magnitude and by spatial distance from the centre of the region of interest. In the

Feature	**Data Intensity Invariances**			
	$+$	$\times$	Mono	Biject
Intensities	·	·	·	·
Intensity rankings (LBPs, ranklets)	★	★	★	·
Autocorrelation-based (MIND, ALOST)	★	★	·	·
Gradients	★	·	·	·
Orientations	★	★	☆	·
Unsigned orientations	★	★	☆	☆
Magnitude-thresholded orientations (orientation histograms)	★	·	·	·
Magnitude-weighted orientations	★	·	·	·
Normalised magnitude-weighted orientations (HOG, SIFT, Toews *et al.* [98])	★	★	·	·

Table 3.1: Intensity transformation invariances of a few feature types. The transformations are (L to R): additive, multiplicative, monotonic and bijective. '★' = invariance, '☆' = semi-invariance and '·' = sensitivity.

case of HOG, simple magnitude weighting is chosen following trials of various weighting schemes as a function of the magnitude. Note that a local normalisation step is first performed and it is the normalised magnitudes which are used, so that these descriptors have some robustness to linear transformation of the data intensities. The HOG descriptor also includes orientation space weighting where each gradient contributes to both bins between whose centres the value of the angle lies, using bilinear interpolation to weight the contributions. We do not employ magnitude weighting since this introduces some dependence on intensity values (at least a linear relationship is assumed between different scan intensity distributions). Rather, voxel contributions are weighted equally and the resulting histogram is normalised, effectively giving a probability distribution. The feature is considered to be missing if less than 50% of voxel values are known. We use an integral volume implementation, similar to that in [83]. Pragmatically, spatial distance weighting would significantly increase run times since we could no longer use efficient integral volume look-ups. However, we do look at the possibility of taking a Gaussian-weighted value in orientation space once the histogram has been computed. This is a similar idea to the blurring that Freeman [79] does using a [1 4 6 4 1] mask in histograms of 36 bins.

The various intensity invariances of different feature types are summarised in Table 3.1. Invariances are an important consideration in selecting features for multi-modal use. For gradient orientations, we describe them as *semi*-invariant since they are only *linearly* invariant within the micro-locality of the pixel i.e. the 4-voxel support over which X and Y gradients are computed by central difference to determine orientation. Compare this to the descriptor of Toews *et al.* [98], which is only linearly invariant within each spatial region over which orientations are binned. If we make the strong assumption that intensities are functions of tissue type, each tissue corresponding to a single intensity value, and further assume spatial coherence, then the invariance to monotonic and bijective intensity transformations holds exactly. Here we define spatial coherence to mean *only two tissues are present* in a micro-locality. See Figure 3.4 for illustration. The validity of these assumptions depends on the voxel resolution and the extent to which noise and texture are absent.

(b) Two-tissue boundary (c) Three-tissue boundary

Figure 3.4: Illustration of the local sensitivity of gradient orientations to nonlinear intensity transforms. Consider the computation of the central pixel's unsigned gradient orientation, $\arctan([y^+ - y^-]/[x^+ - x^-])$. a) The orientation is identical for A, B and C. b) The orientation is identical for A, B and C which are linearly related, but different for D which is nonlinearly transformed. Hence, the presence of a three-tissue boundary (which does not satisfy our spatial coherence criterion) would not guarantee invariance to monotonic or bijective (in the case of *unsigned* orientations) transformations.

3.2.4 Feature notation

We define our feature f_{orient} as follows.

$$f_{orient}(b, v, d, C, \psi) = P(b|v + d, C, \psi) \tag{3.1}$$

This new gradient orientation feature f_{orient} is the relative frequency of a voxel gradient in the plane ψ ($\psi \in$ {axial, coronal, sagittal}) being oriented within the angle range of the ith bin b ($b = 1...B$ bins), in the cuboid of dimensions C = {C_x,C_y,C_z} lying at an offset d from the voxel of interest v. In the next section, results are presented from the optimisation of these feature parameters.

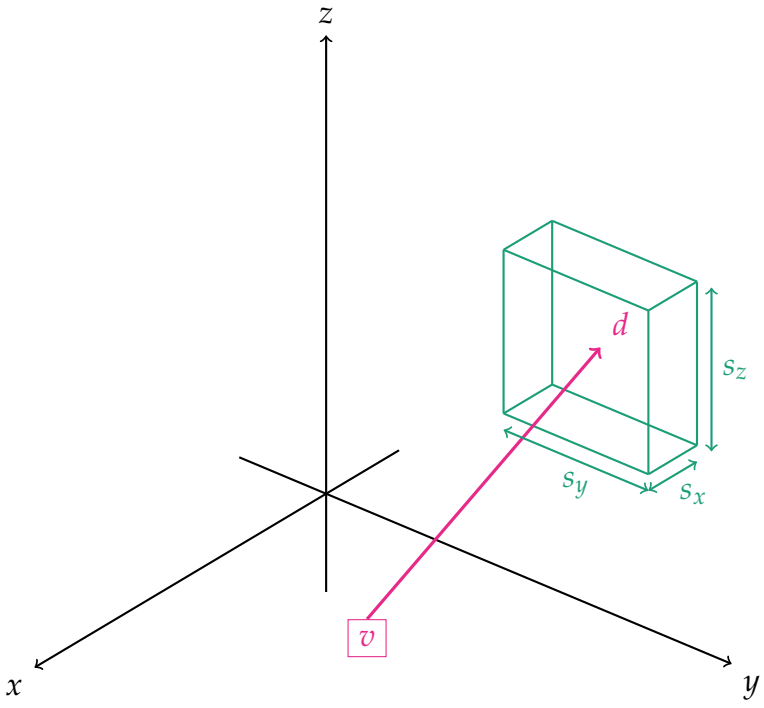


Figure 3.5: Illustration of the spatial parameters of the gradient orientation feature. The histogram for any given feature is computed from the intensity values in the cuboid of dimensions C = {C_x,C_y,C_z} at offset d from voxel v.

In section 3.4, we start by exploring the basic feature parameter space, specifically the dimensions of the cuboids over which the histogram is computed, the number of features selected per tree, the pattern of random offset sampling for the cuboids, the resolution of the histogram and the plane in which the 2D gradient orientations are measured. We arrive at a set of optimised parameters, in the process obtaining some information as to how sensitive the detector is to each parameter.

Following optimisation of the feature parameters, in section 3.5 the detector is validated on a separate test cohort, where the trade-offs between accuracy, speed and memory usage are illustrated. In section 3.6 gradient orientation features are applied to the CT problem of chapter 2 to see if they improve performance over intensity features alone. Finally in section 3.7, unsigned gradient orientations are employed to train and cross-validate a cross-modality detector (i.e. the detector is trained on one modality and validated on another), using MRI-T1, MRI-T2 and CT data cohorts.

The term *HOG* is not used to describe our features because the feature formulation that we arrive at does not contain the refinements which distinguish HOG features from SIFT features and the orientation histograms of Freeman *et al.* [79, 81]. Rather, the generic term *gradient orientation features* is used to refer to the features that we develop through this chapter.

3.3 Data

Figures 3.6 and 3.7 show demographics of the data that we use to optimise and validate gradient orientation features.

- The 50 sagittal MRI-T1 training datasets and the 35 axial MR-T1 datasets will be used to optimise the f_{orient} parameters.
- The 24 sagittal MRI-T1 test datasets will be used at the end for validation.
- The 60 MRI-T2 axial datasets will be used in a validation experiment to observe the effectiveness of f_{orient} features for an inter-modality detector.

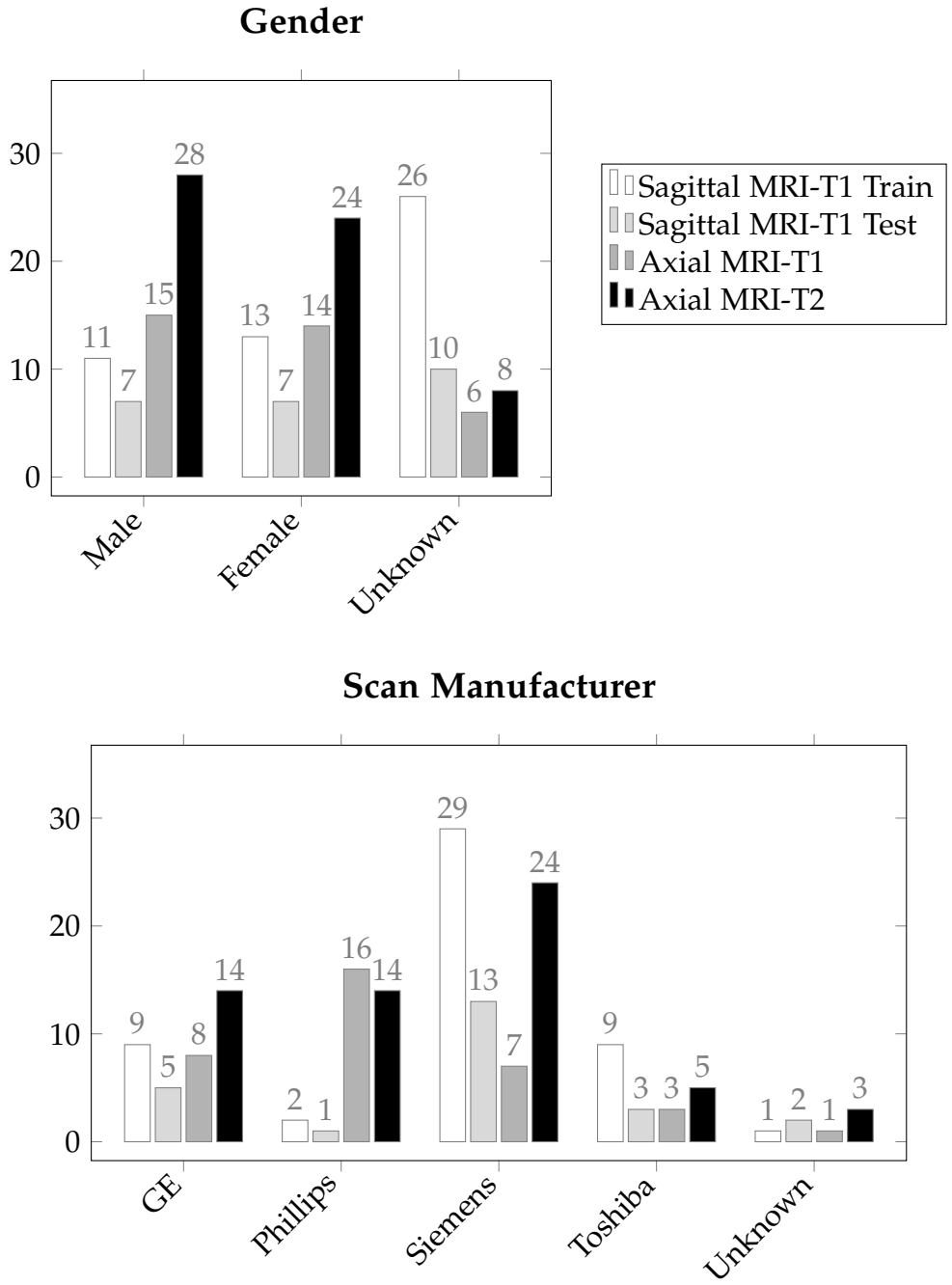
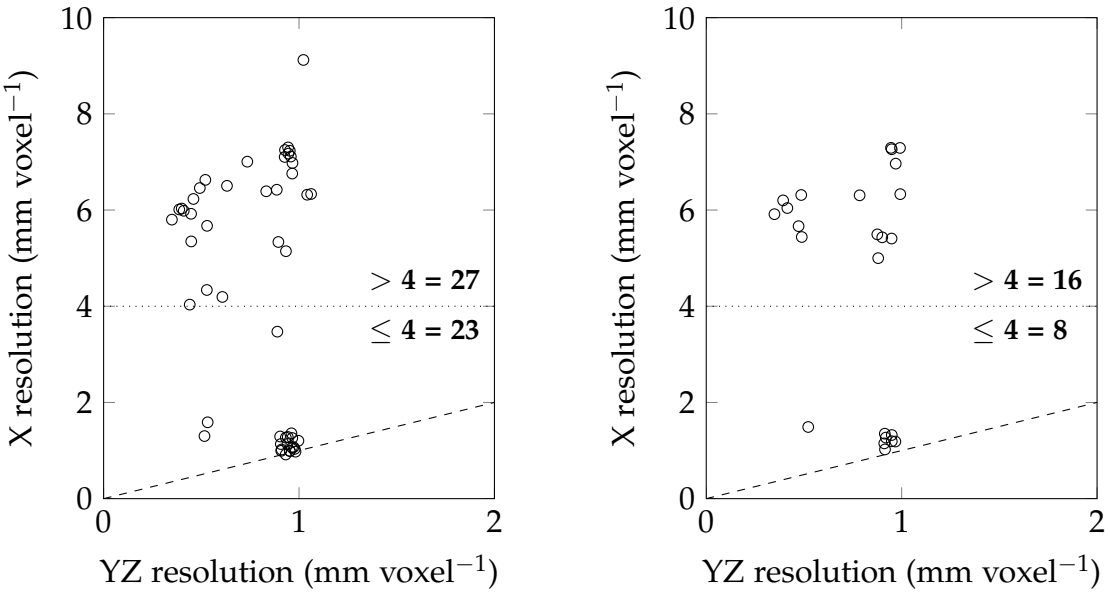
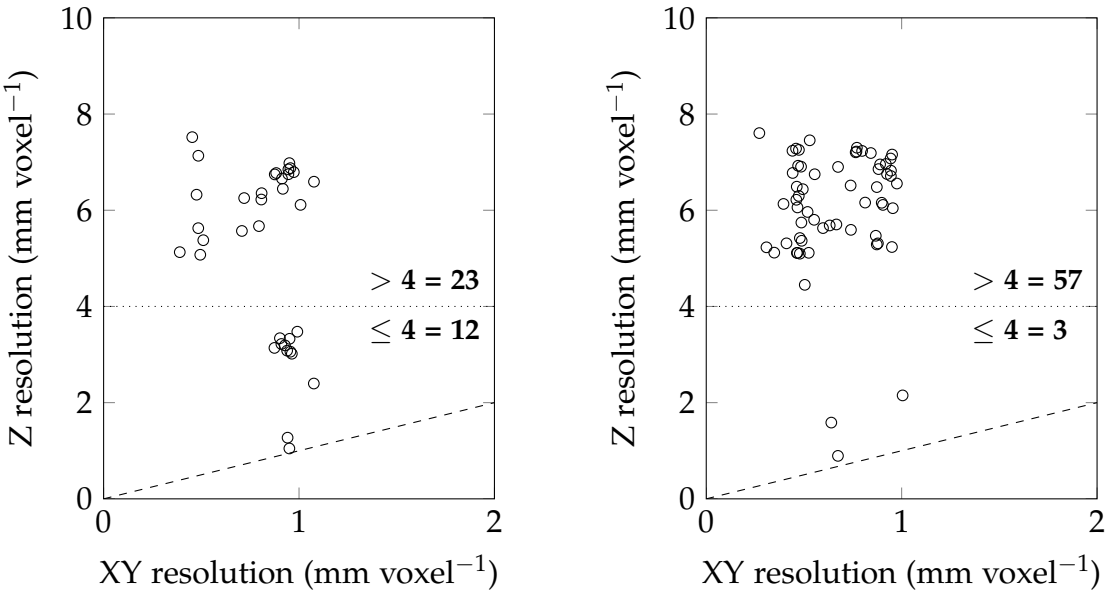


Figure 3.6: Plots showing the distribution of gender and scan manufacturer for the four cohorts of data which will be used to train and validate gradient orientation features in this chapter.



(a) Sagittal MRI-T1 Train

(b) Sagittal MRI-T1 Test



(c) Axial MRI-T1

(d) Axial MRI-T2

Figure 3.7: These (jittered) graphs show the resolution values which are represented in the four cohorts of data which will be used to train and validate gradient orientation features in this chapter. Jitter has been added to the points because there are cases where many datasets have the same resolution. Voxels are isotropic in the plane of the acquisition direction (horizontal axis), but the slice thickness (vertical axis) varies. The dashed line on these graphs indicates isotropy. The writing above and below the dotted line at 4mm voxel^{-1} gives the exact figures of datasets with slice thickness above and below the 4mm voxel^{-1} scale which is employed during landmark detection.

3.4 Exploration of gradient orientation features

In this section, aspects of gradient orientation features are investigated to determine the best configuration for landmark detection.

3.4.1 Experiment procedure

Methodology

Random forest detector: For these experiments, a pared down detector is trained with $T = 40$ and $D_T = 8$. We use the same parameter values as chapter 2 as follows: $\sigma_{Sampling} = 3.0$mm, $B_{Ratio} = 5.0$, $w_{Node\_min} = 5.0$, $w_{Node\_Split\_min} = 2.0$ and $d_{skip} = 2$ voxels.

Data: The sagittal MRI-T1 training dataset cohort is used to train and test the detector for the majority of experiments, except for the experiments investigating the feature plane where the axial MRI-T1 cohort is also used. An out-of-bag validation strategy is employed (see section 2.3.4).

Number of Feedback Iterations: Experiments are generally run for the zeroth iteration except for inconclusive cases, where atlas location autocontext is employed (using an *affine* transformation) and the first iteration is also run.

Reporting of results

Metrics: The mean landmark error and AUC (for a 30mm Localisation Receiver Operating Curve (LROC)) are reported. Additionally the *capped* mean landmark error is given (indicated with a dotted line on the graphs), computed from errors which are capped at 30mm; this is included to show results without the distortions arising from large errors.

Repetitions: Three runs are performed for each experiment, using a different random seed for each, to give an idea of the inherent noise in the method due to randomness. The graphs show results from all runs.

Graphs: Each run comprises a sweep of the feature parameter under experimentation. In the experiment graph, a line is drawn through all results in a run since all are acquired using the same random selection of training data samples (according to the run's random seed) — but the feature selection is different.

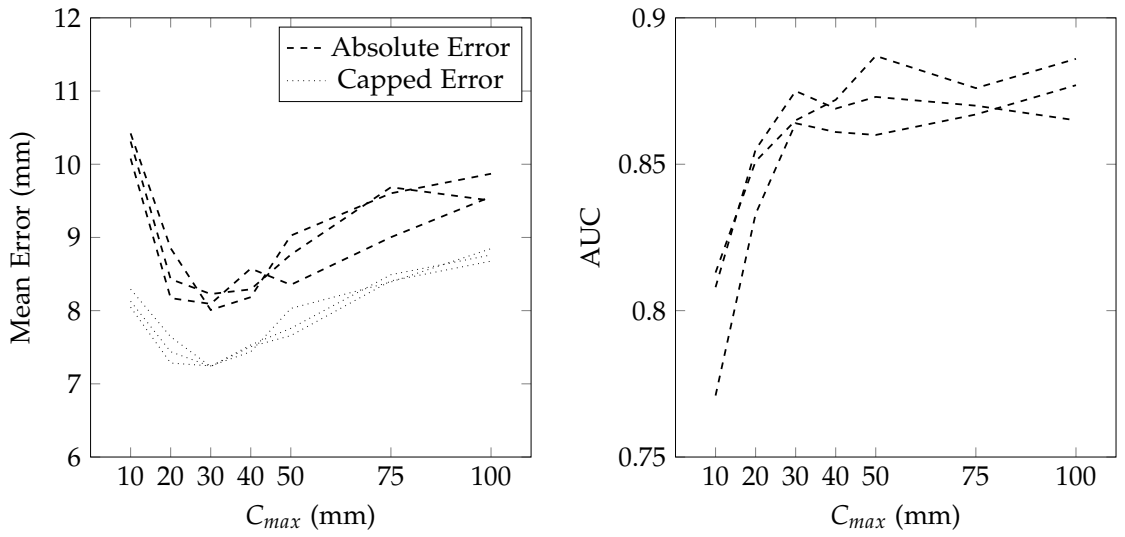
3.4.2 Size of feature cuboid

Description: The histogram for each feature is computed over a cuboid. The dimensions $C = C_x, C_y, C_z$ of the cuboid are randomly selected, from the range {1,C_{max}}. In this experiment, we investigate the optimal size limit C_{max}.

Method: A random forest detector is trained on the sagittal MRI-T1 training cohort using B = 8, F_T = 2500, d_{max} = 52mm, ψ = *sagittal* and volumetric feature sampling. Cuboid sizes C_{max} = {10, 20, 30, 40, 50, 75, 100mm} are trialled.

Results: Figure 3.8 indicates that the optimum value of C_{max} is 30mm.

Discussion: There is a trade-off between spatial resolution and noise sensitivity. Very large cuboids lack spatial resolution and give rise to missing values for many voxels (since the information is considered to be missing if more than 50% of voxels inside the cuboid are missing). Very small cuboids capture little information, and will be sensitive to noise.



(a) Mean Error (b) AUC

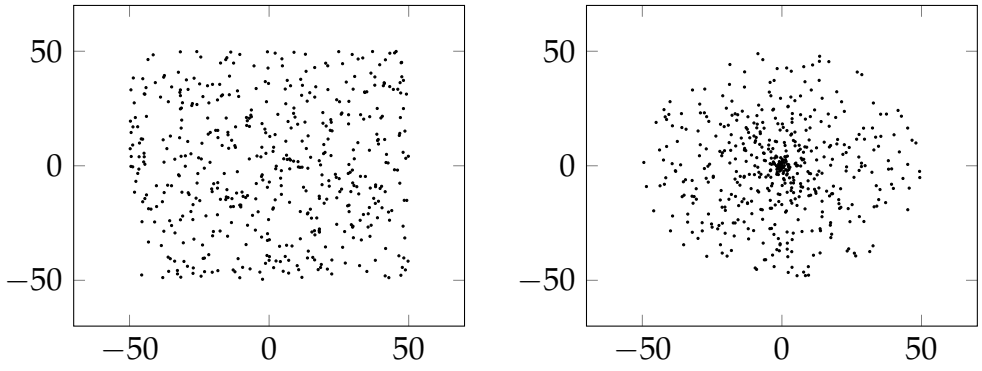
Figure 3.8: Graphs showing the effect of maximum feature cuboid size C_{max} on landmark detection accuracy. Cuboid sizes of C_{max} = {10, 20, 30, 40, 50, 75, 100mm} are trialled. It can be seen that for these parameters, C_{max} = 30mm is optimum.

3.4.3 Number of features and feature sampling strategy

Description: This experiment is run to find the best feature sampling strategy. We test two strategies ("radial" and "volumetric"), using varying numbers of features for each.

- *Radial* sampling refers to uniform sampling with respect to offset magnitude.
- *Volumetric* sampling refers to uniform sampling with respect to volume i.e. uniform density.

Figure 3.9 illustrates the appearance of the sampling distributions in two dimensions.



(a) Volumetric Sampling (b) Radial Sampling

Figure 3.9: Illustration of the volumetric and radial sampling strategies. The points represent sampling positions relative to the voxel of interest at (0,0). The axis units are mm. These diagrams are in 2D but we are actually working in 3D space, so the volumetric method samples a cuboid (rather than a square) and the radial method samples a sphere (rather than a circle).

Method: A random forest detector is trained on the sagittal MRI-T1 training cohort using $B = 8$, $C_{max} = 30$mm, $d_{max} = 52$mm and $\psi = sagittal$. Feature numbers of $F_T = \{500, 1000, 2500, 5000\}$ are trialled.

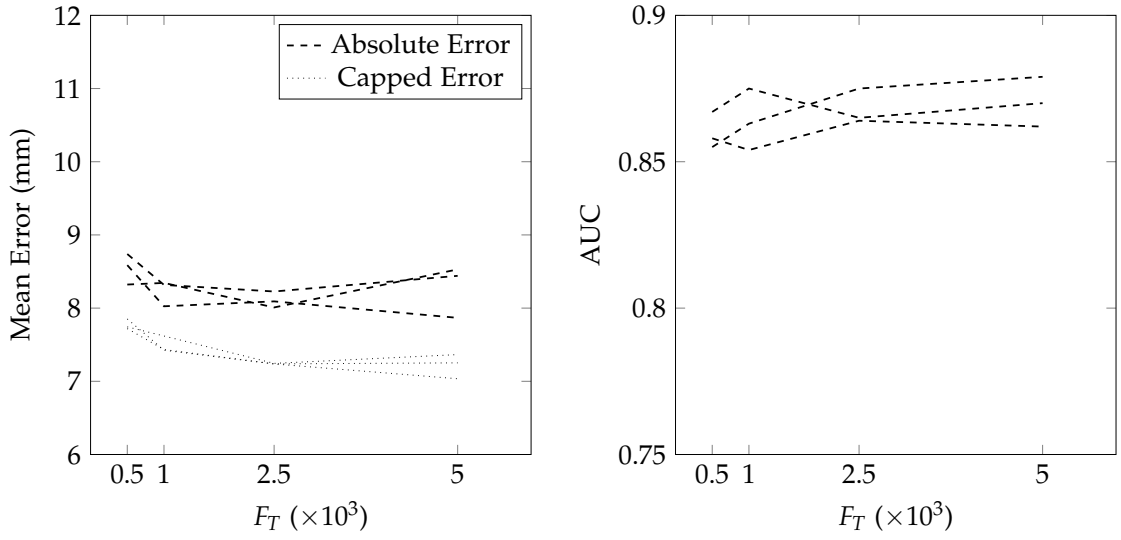
Results: See the results in Figures 3.10 (zeroth pass) and 3.11 (first pass). The optimum number of features is 2500, although any number within the range of 1000 – 5000 appears to give reasonable results.

In terms of sampling strategy, in the zeroth iteration radial sampling gives a better capped landmark error, but poorer absolute landmark error, than volumetric sampling. When atlas location autocontext is deployed, better results

are achieved by those combinations which use radial sampling in iteration 1. After iteration 1 there is little difference between the capped and absolute results, which reflects the fact that atlas location autocontext is correcting gross outliers.

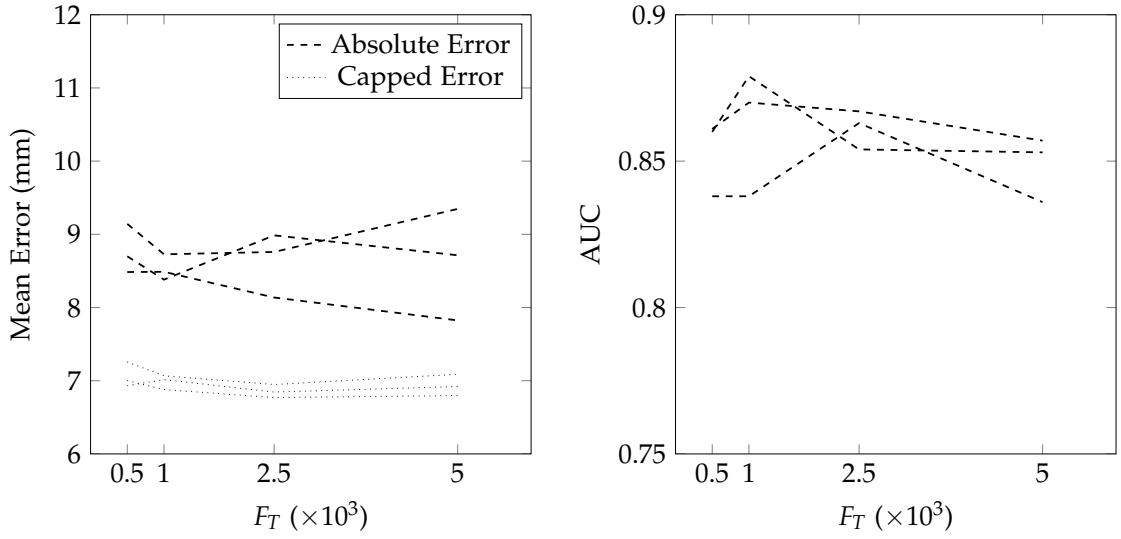
Discussion: Too few features gives a worse result due to a paucity of information. Too many features gives a worse result due to the reduction in randomisation: at the limit, every tree would be given the whole set of features, and the trees would be identical, were it not for the randomised selection of training samples and bagging of datasets.

Radial sampling places greater emphasis on local appearance and less on wider spatial context, compared to volumetric sampling. As a result, landmarks are more precisely placed but are more likely to be wildly wrong without atlas location feedback. See Figure 3.12 for images from an example dataset which illustrates the effect of atlas feedback.



(a) Volumetric Sampling, Mean Error

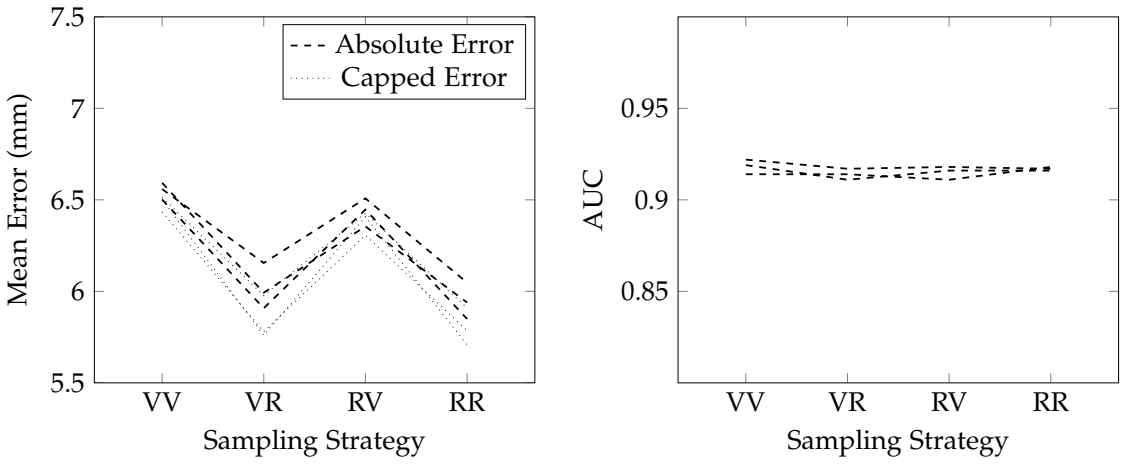
(b) Volumetric Sampling, AUC



(c) Radial Sampling, Mean Error

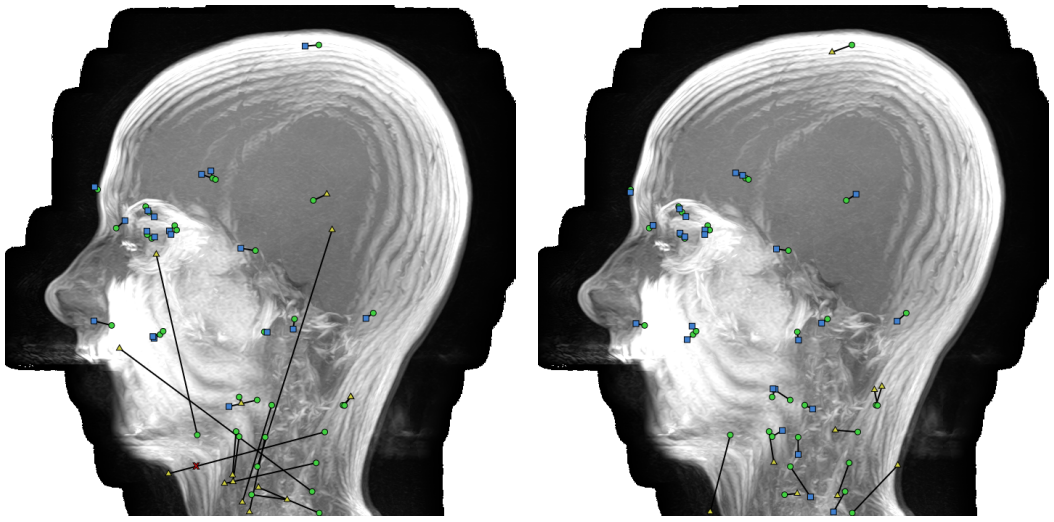
(d) Radial Sampling, AUC

Figure 3.10: Graphs showing the effect of the number of features selected per tree, F_T, on landmark detection accuracy after iteration 0. Above: Volumetric sampling. Below: Radial sampling. Values of F_T = {500, 1000, 2500, 5000} are trialled.



(a) Mean Error (b) AUC

Figure 3.11: Graphs showing the effect of feature sampling strategies on landmark detection accuracy over two iterations. Each combination of the two sampling strategies is tried for the two iterations. "V" stands for volumetric and "R" stands for radial so e.g. RV indicates radial sampling in the zeroth iteration, and volumetric sampling in the first iteration.



(a) Iteration 0 (b) Iteration 1

Figure 3.12: Sagittal MIP images of an example dataset showing the effect of atlas location autocontext. These results were obtained using *radial* feature sampling. There is particularly poor performance for the landmarks in the neck (cervical vertebrae, carotid arteries etc.). The errors are much reduced following atlas coordinate feedback. This dataset also illustrates the close cropping of datasets which is present in many MRI datasets. Note that the striping is an artefact of the MIP visualisation which arises as a result of the large slice spacing. [TMVS Dataset ID: 4729]

3.4.4 Histogram resolution and Gaussian windowing

Description: In this section, we look at what histogram resolution gives the best result. We also experiment with Gaussian windowing (see Figure 3.13), using standard deviations up to 45 degrees in size.

Method: A random forest detector is trained on the sagittal MRI-T1 training cohort using F_T = 2500, radial feature sampling, C_{max} = 30mm, d_{max} = 52mm and ψ = *sagittal*. Values of B = {2,4,8,16} are trialled.

In a follow-up experiment, Gaussian windowing is tested. For B = 8, Gaussian windows are tried with standard deviations of 0.7 and 1.0 bin widths i.e. 31.5 and 45 degrees. For B = 16, Gaussian windows are tried with standard deviations of {0.7, 1.0, 1.5, 2.0} bin widths i.e. 16, 23, 34 and 45 degrees.

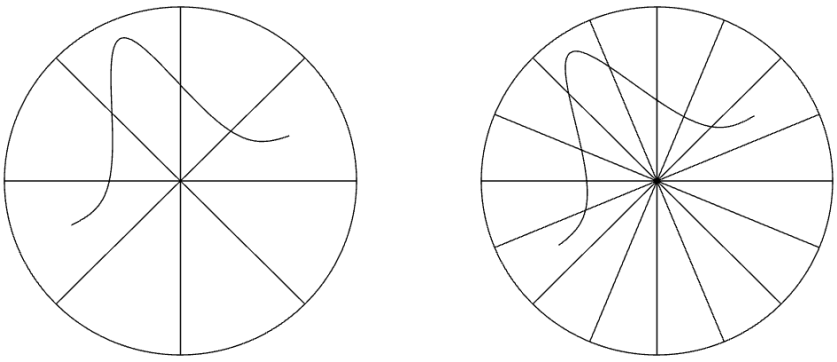


Figure 3.13: Illustration of Gaussian windowing. The sectors of the circles represent the bins of the histogram (8 bins in the left-hand picture and 16 bins in the right-hand picture). When Gaussian windowing is employed, weights are assigned to the bin of interest and its neighbours, according to a Gaussian function with standard deviation σ centred at the bin of interest, and a weighted bin value is computed. Due to the cyclical nature of the distribution, if a large enough sigma was used, there would be a wrap-around effect.

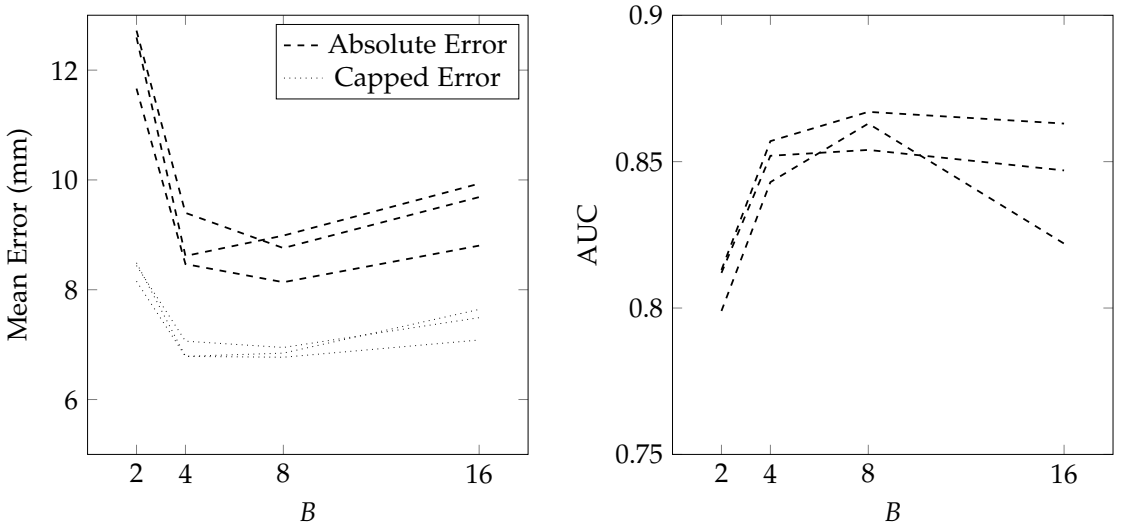
Results: Results are shown in Figures 3.14 and 3.15. The optimal B is 8 (out of 2, 4, 8 or 16 bins) although we could halve the number of bins to 4 with only a small decrease in accuracy. This is a useful possibility since in an integral volume implementation, less memory is required for fewer bins. Using a Gaussian mask improves the result, particularly for the 16-bin histogram features, which then give

as good (or possibly better) performance than 8 bins. However, the improvement over simply using 8 bins without Gaussian weighting is not significant.

Discussion: Too few or too many bins and the shape of the histogram is lost. Fine division of orientations also makes the descriptor highly sensitive to the variation in patient posture which is inevitably present (in particular, rotation of the neck and jaw). This is the same bin size as used in the SIFT descriptor [88, 90], but a larger bin size than was used in the HOG descriptor (9 bins over 180 degrees).

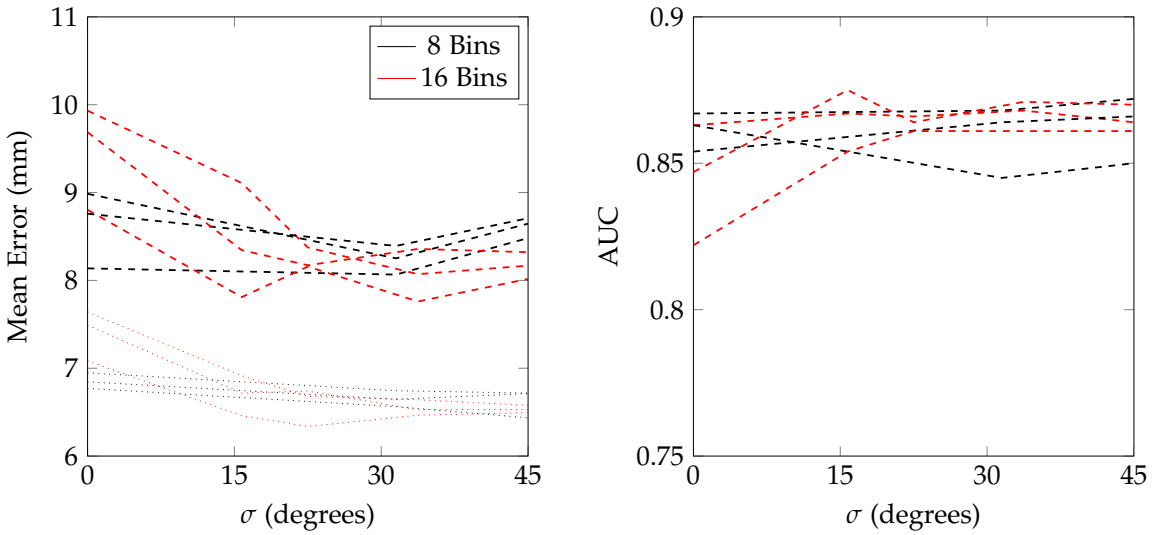
Gaussian windowing will be omitted in future experiments since the improvement is not convincing enough to merit the extra volume access operations and extra complexity. Normally 16 integral volume access operations are required, to read values from two integral volumes: that of the bin of interest and that of the "sum of known values" (the latter required for histogram normalisation, and also to verify if $\geq 50\%$ of values are known). When Gaussian windowing is used, a further 8 volume accesses are required for every extra bin included in the weighted computation.

Additional experiments were done, varying the positioning of the bins (i.e. instead of separators at {0, 45, 90 ...} with respect to volume space, use separators at {-22.5, 22.5, 67.5 ...}) and also using bilinear interpolation when assigning orientations to bins (i.e. each orientation contributing to both bins between whose centres the value lies, as is done in the HOG descriptor). These changes to the metholodology produce no significant difference and are not considered further.



(a) Mean Error (b) AUC

Figure 3.14: Graphs showing the effect of the number of bins B on landmark detection accuracy. Values of B = {2,4,8,16} are trialled.



(a) Mean Error (b) AUC

Figure 3.15: Graphs showing the effect of Gaussian windowing on landmark detection accuracy. For an 8-bin histogram (black) standard deviations of 31.5 and 45 degrees (i.e. 0.7 and 1.0 bin widths) are used, and for 16 bins (red) standard deviations of {15.75, 22.5, 33.75, 45} degrees (i.e. 0.7, 1.0, 1.5 and 2.0 bin widths) are used.

3.4.5 Plane of feature and plane of scan acquisition

Description: In this section, gradient orientations are measured in different planes to see which gives the best result. The hypothesis is that gradients will be most reliably measured in the plane in which the volume was acquired.

Method: To test this hypothesis, experiments are run for both sagittal MRI-T1 and axial MRI-T1 populations. For each data cohort, a detector is trained with different planes ψ:

- Axial features only
- Sagittal features only
- Features from all planes (sagittal, axial and coronal)
- Features from all planes but only one plane used per tree

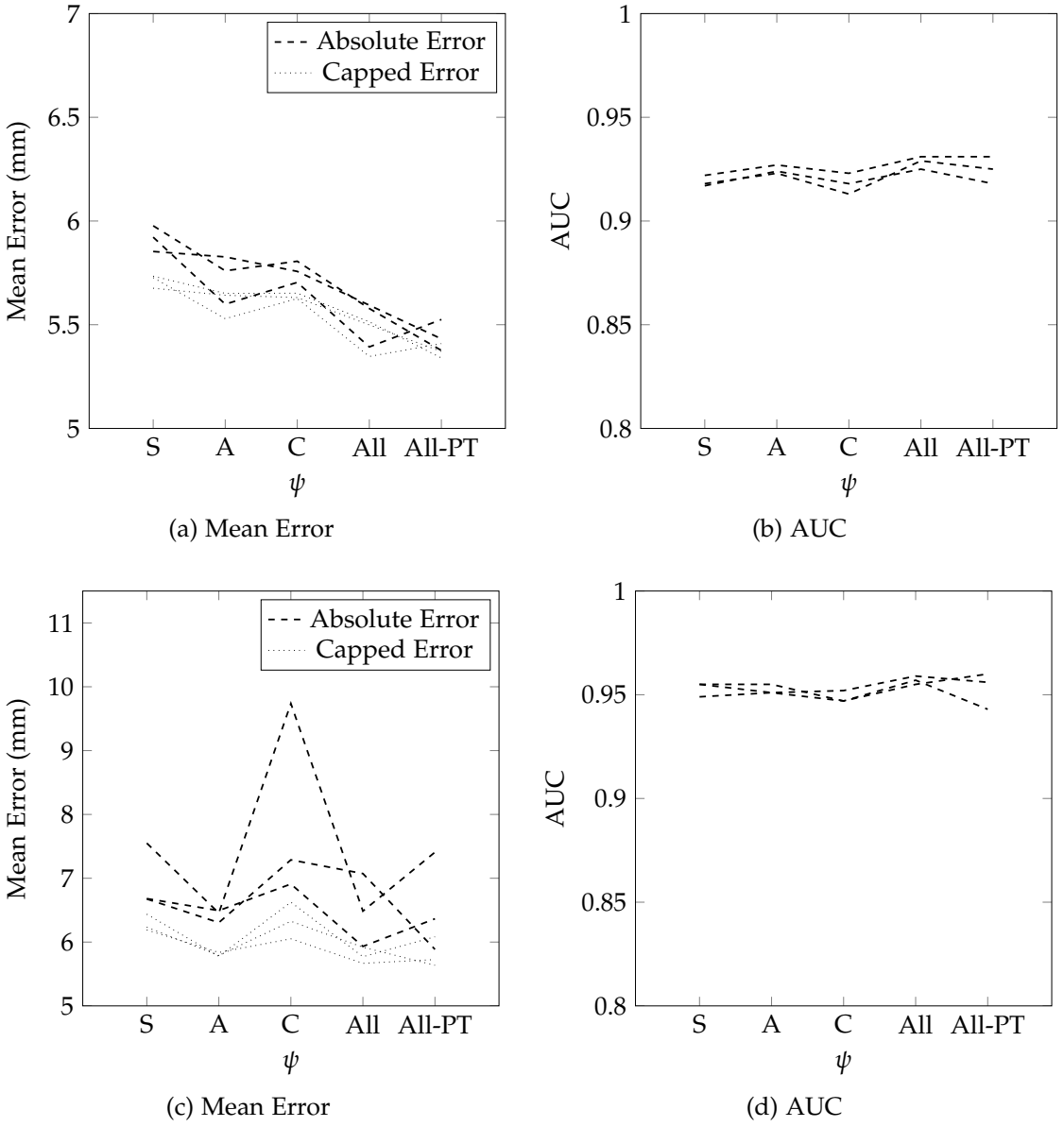
The parameters used are F_T = 2500, radial feature sampling, B = 8, C_{max} = 30mm and d_{max} = 52mm. The experiments are run at $D_{Res} = 4\text{mm voxel}^{-1}$ and $D_{Res} = 1\text{mm voxel}^{-1}$. For the latter experiment, we use $\sigma_{Sampling}$ = 0.75mm in order to to yield the same number of samples as at the lower resolution.

Results: Results are given after iteration 1 in Figure 3.16 (4mm voxel^{-1}) and Figure 3.17 (1mm voxel^{-1}). Unexpectedly, at the standard scale of 4mm voxel^{-1} the axial features give best performance for both data cohorts. The larger slice spacing does not appear to compromise the reliability of features measured in planes orthogonal to the plane of acquisition. The experiment was also run at the higher resolution of 1mm voxel^{-1}. In this case, the plane of acquisition did correlate with the best feature plane. Figure 3.18 illustrates why; the slice spacing is much more obvious at the higher resolution.

Conclusion: At 1mm voxel^{-1}, the plane of acquisition correlates with the best performing feature plane. At the standard detection resolution of 4mm voxel^{-1} (which appears to give little regression in accuracy), the plane of acquisition makes no significant difference. Rather, axial plane features perform best for both the sagitally and axially acquired data.

Discussion: It is an interesting observation that at a resolution of 4mm voxel^{-1}, the axial plane features appear to give best performance no matter the plane of acquisition. To check that the sagittal displacement feature is not muddying

the waters, the experiment was repeated without using the sagittal displacement feature. All results were a little worse, but axial features still outperformed sagittal features (results not shown here). For the axially acquired data, the axial features alone appear to perform better than using features from all planes. This suggests that the nature of the landmarks, the data, or the human body structure lends itself to better detection in the axial plane. This finding merits further investigation with more data, including scans of other body parts.



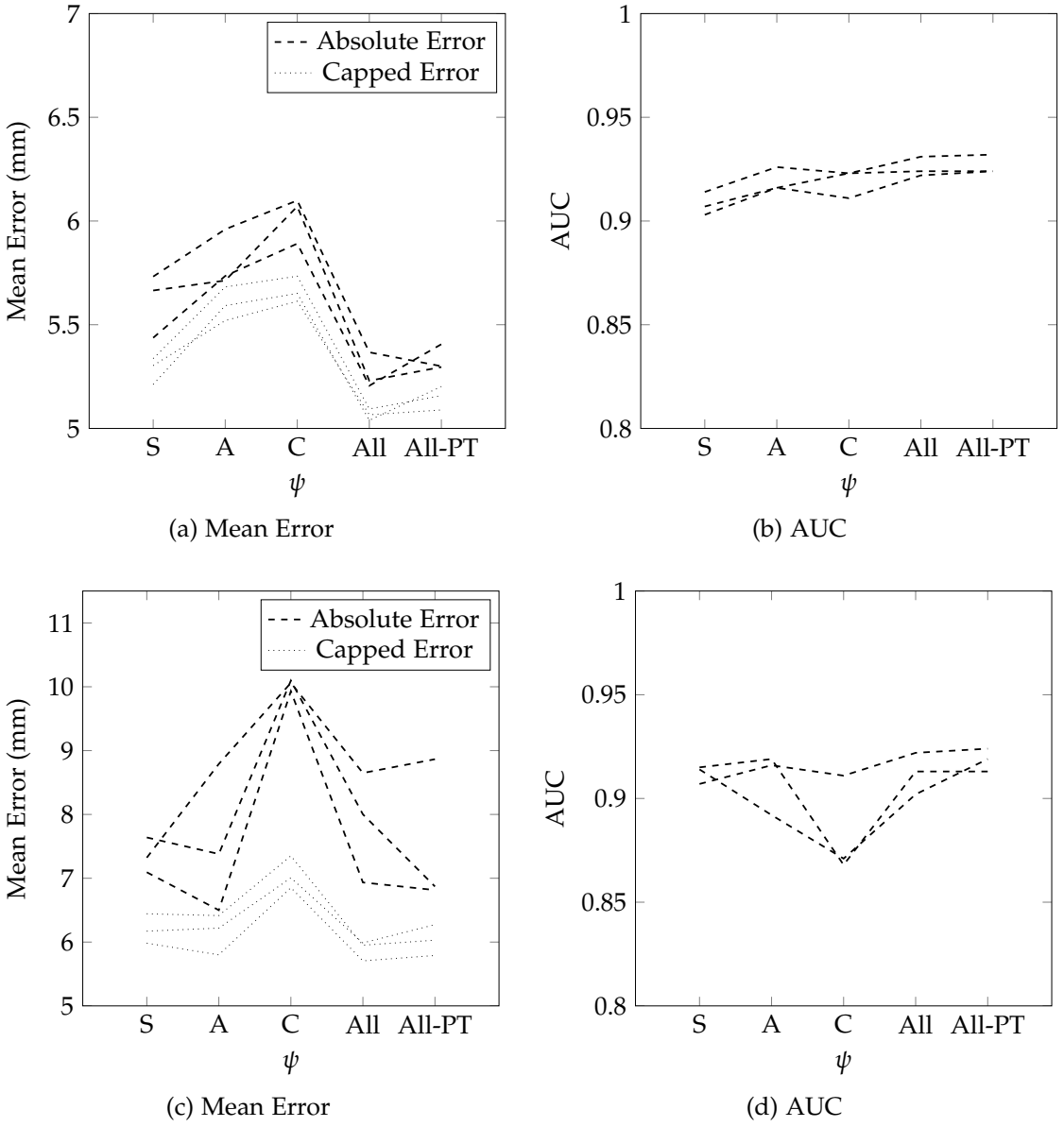
(a) Mean Error

(b) AUC

(c) Mean Error

(d) AUC

Figure 3.16: Graphs showing the effect of feature plane ψ on landmark detection accuracy. Results are given after iteration 1, for detectors trained on datasets scaled at 4mm voxel^{-1} resolution. Above: Sagittal MRI-T1 training cohort. Below: Axial MRI-T1 cohort. S = Sagittal features, A = Axial features, C = Coronal features, All = Sagittal/Axial/Coronal features, All-PT = Sagittal/Axial/Coronal features (*One plane per tree*).



(a) Mean Error

(b) AUC

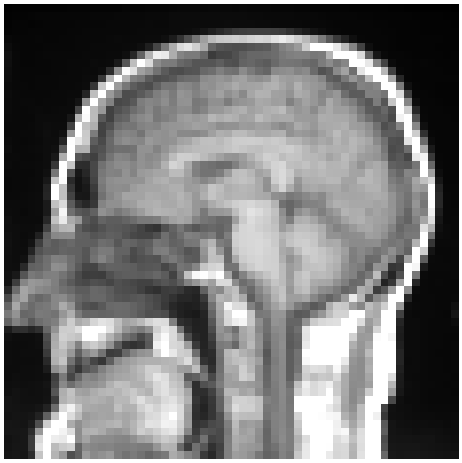
(c) Mean Error

(d) AUC

Figure 3.17: Graphs showing the effect of feature plane ψ on landmark detection accuracy. Results are given after iteration 1, for detectors trained on datasets scaled at 1mm voxel^{-1} resolution. Above: Sagittal MRI-T1 training cohort. Below: Axial MRI-T1 cohort. S = Sagittal features, A = Axial features, C = Coronal features, All = Sagittal/Axial/Coronal features, All-PT = Sagittal/Axial/Coronal features (*One plane per tree*).

(a) Mid-sagittal slice, Full resolution

(b) Mid-coronal slice, Full resolution



(c) Mid-sagittal slice, 4mm voxel^{-1}

(d) Mid-coronal slice, 4mm voxel^{-1}

Figure 3.18: Mid-volume slices from an example sagittally acquired MRI-T1 dataset with a resolution of 0.9x0.9mm voxel^{-1} in the sagittal plane, and slice spacing of 6.96mm. The large slice spacing is much less obvious at the detection resolution of 4mm voxel^{-1}. [TMVS Dataset ID: 4664]

3.4.6 Noise detection and thresholding

Description: Many of the gradients being detected are not reflective of true structural detail, but are small perturbations in homogenous regions which arise from image noise. This is particularly obvious in regions of air. In this section, we look at the possibility of determining a noise threshold and binning gradients with sub-threshold gradients in a designated zero-gradient bin.

Method: The gradient magnitudes are first made approximately equivalent by linearly normalising volume intensities such that the 5th and 95th voxel intensity percentiles map to 0 and 1000 respectively. Extreme values are allowed to fall outside of this range (rather than clamping to 0 and 1000). Figure 3.19 shows the same image slice as shown in Figures 3.2 and 3.3, thresholded at different gradient magnitudes.

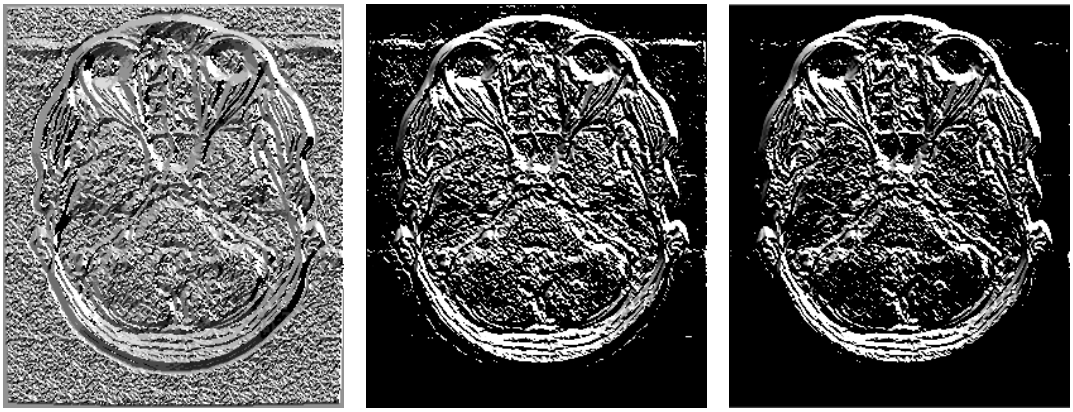
A random forest detector is trained on the sagittal MRI-T1 training data cohort, using F_T = 2500, radial feature sampling, B = 8, C_{max} = 30mm and d_{max} = 52mm. Features from all planes are used (one plane per tree).

Results: The results are shown in 3.20.

Conclusion: There is no obvious benefit to using a noise threshold, at least with the threshold-finding method used here.

Discussion: Where noise has a truly random distribution and the histogram is computed over a sufficiently large region, the resulting histogram shape will be uniform (see Figure 3.3). These experiments suggest that trusting to the random distribution of noise is sufficient. Further, use of a noise threshold weakens the invariance to (monotonically related) gradient magnitudes. The sacrifice of useful invariances may explain why a noise threshold is not beneficial overall.

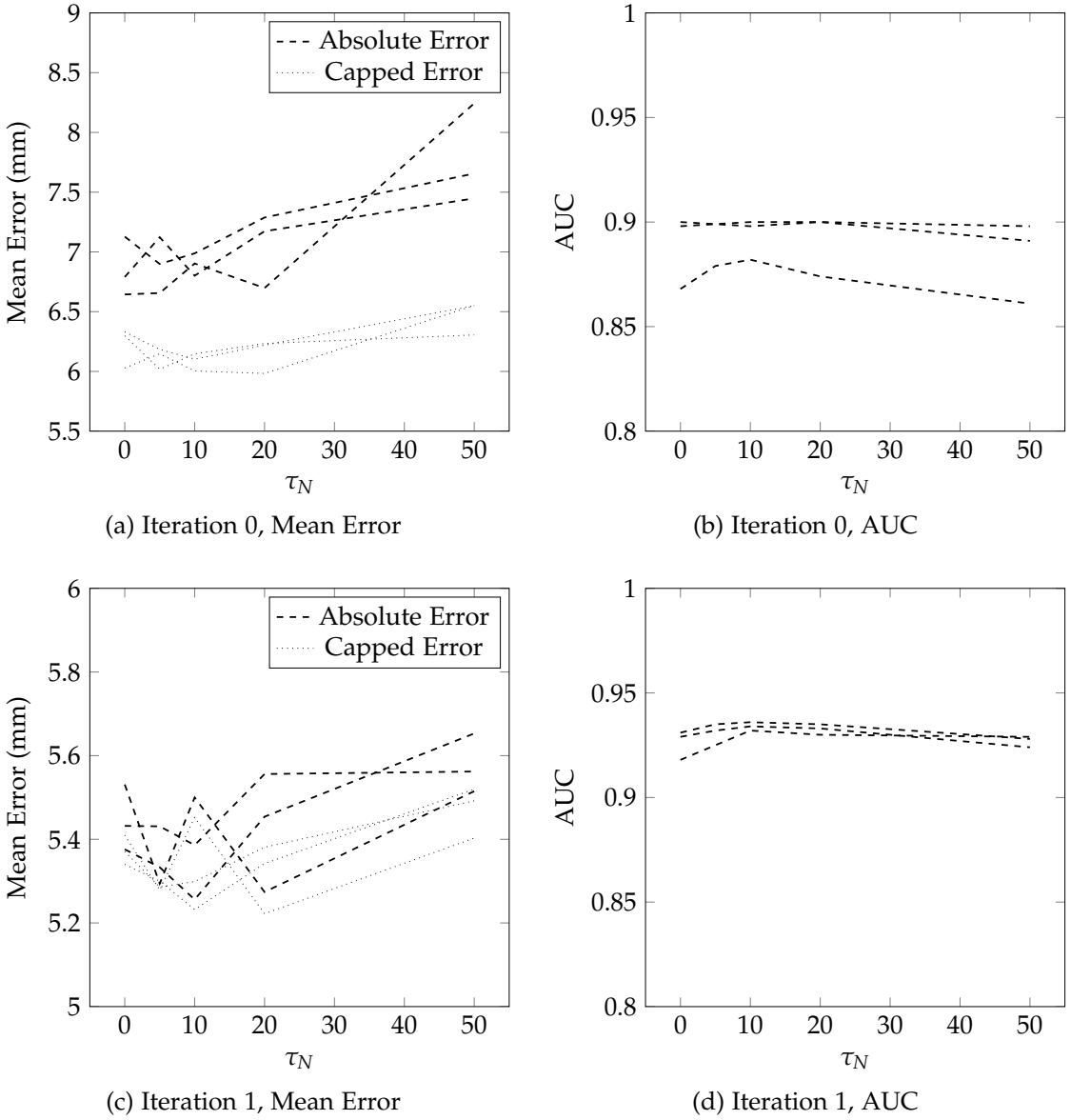
We choose not to employ a noise threshold in later experiments. However, we predict that minor artefacts such as ringing, striping or bias field gradients might cause problems. A subject for future investigation is investigation of methods for automatic noise threshold-finding.



(a) Gradient Image (b) τ_N = 5 (c) τ_N = 10

(d) τ_N = 20 (e) τ_N = 50 (f) τ_N = 100

Figure 3.19: Images of the gradients in an example axial slice showing the effect of applying different noise level thresholds. Gradients with a magnitude below the noise level threshold (τ_N) are shown in black. [TMVS Dataset ID: 4660]



(a) Iteration 0, Mean Error

(b) Iteration 0, AUC

(c) Iteration 1, Mean Error

(d) Iteration 1, AUC

Figure 3.20: Graphs showing the effect of applying a noise level threshold τ_N on landmark detection accuracy. Values of τ_N = {0,5,10,20,50} are trialled. The results from iteration 0 are shown above, and those from iteration 1 are shown below.

3.4.7 Mixing gradient orientation and intensity features

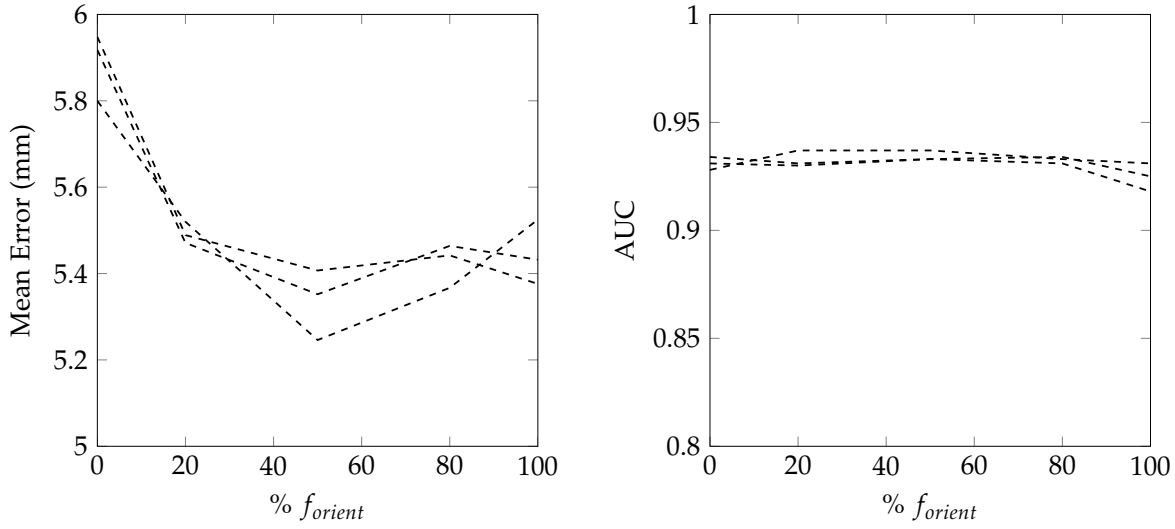
Description: In this section, gradient orientation features are compared with intensity features. The question is whether, if the modality is known (for instance MRI-T1), is it more effective to train a detector using intensity features or using gradient orientation features, or perhaps a mixture of both? It is expected that the answer will be different for CT than for uncalibrated modalities. Hence in this section we run two experiments, one on the sagittal MRI-T1 cohort and one on the CT head data from the training cohort of chapter 2.

Method: A series of experiments are run in which the ratio of gradient orientation features to intensity features is varied (keeping F_T = 2500). As mentioned previously, the data is pre-processed by linearly normalising volume intensities such that the 5th and 95th voxel intensity percentiles map to 0 and 1000 respectively. Relative intensity features are then used i.e. the intensity at an offset d minus that of the voxel itself, $I(v + d) - I(v)$. The random forest detector is trained on the sagittal MRI-T1 training data cohort, using radial feature sampling, B = 8, C_{max} 30mm, d_{max} = 52mm and ψ = {*axial, coronal, sagittal*} (one plane per tree).

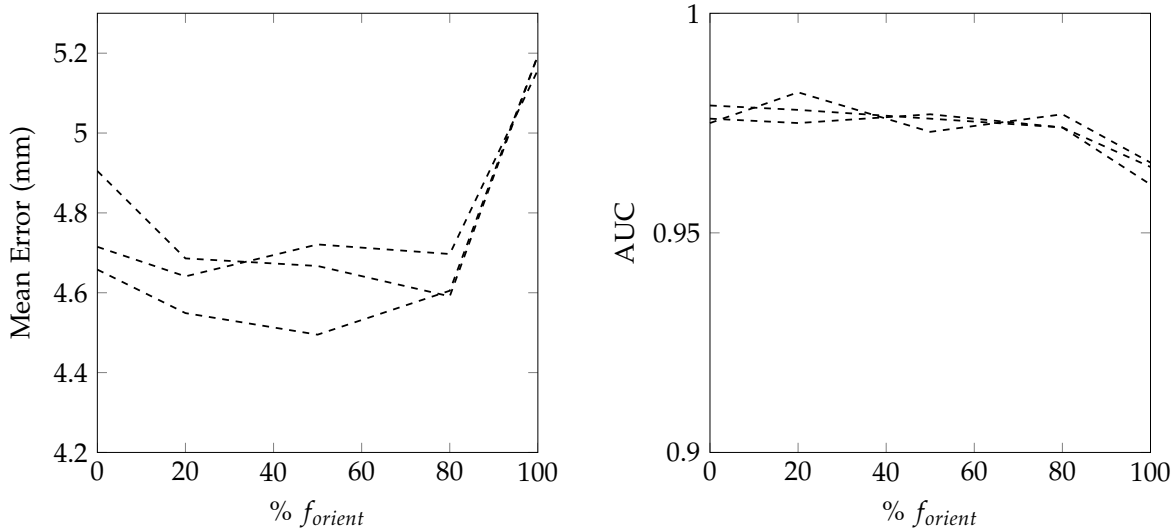
Result: Results are shown in Figure 3.21.

Conclusion: Gradient orientation features perform better than intensity features in MRI, whereas the reverse is true in CT. A 50-50 mix of features gives the optimal result in CT. There may be some benefit to mixing features in MR, however more data is needed to show this.

Discussion: As expected, the feature which performs best changes with modality. The intensity features do not show a convincing additive benefit to using only gradient orientation features in MRI. Even when relative intensities are used, on linearly normalised data, there is still some dependence on the shape of the original data distribution.



(a) Mean Error (b) AUC



(c) Mean Error (d) AUC

Figure 3.21: Graphs comparing the accuracy of intensity features and gradient orientation features, for landmark detection in head MRI (top) and CT (bottom) data. Results are given after iteration 1. 2500 features per tree are chosen for each detector. The horizontal axis shows the percentage of features which are gradient orientation features. For instance, 20% denotes 500 gradient orientation features and 2000 intensity features.

3.4.8 Conclusions

In summary, we find that:

- A maximum box size C_{max} of 30mm is optimum.
- A radial pattern of sampling (uniform sampling with respect to offset magnitude) gives better results than volumetric sampling (sampling with uniform density over the local neighbourhood).
- The detector is fairly robust to the number of features F_T selected per tree: any number within the range of 1000 – 5000 appears to give reasonable results.
- The optimal number of histogram bins B is 8, although there is little degradation when downsizing to 4 bins, which is a useful result where memory usage is a concern. Using a mask improves the result, particularly for the 16-bin histogram features, which then give as good, or possibly better, performance than 8 bins. The improvement over simply using 8 bins without Gaussian weighting is not significant.
- Axial plane features appear to perform better than sagittal or coronal features. Further investigation is required to determine why. Where we move to detection at a resolution which is significantly higher than the slice resolution, then features in the plane of acquisition become most accurate.
- Utilisation of a noise level threshold has no obvious benefit.
- Gradient orientation features perform better than relative intensity features in MR images. The reverse is true in CT images. A 50-50 mix appears to give a slight improvement in the case of CT.

Note that these conclusions hold for a resolution of 4mm voxel^{-1}, and that a different set of parameters will likely best performance at another resolution. For instance, the plane of acquisition was shown to matter more when operating at a resolution (1mm voxel^{-1}) that was at the high end of the slice spacing range.

3.5 Performance characteristics of a gradient orientation detector

There are a number of trade-offs and constraints to take into consideration when choosing random forest parameters. These include:

- Accuracy
- Detection run time
- Training run time
- Size of required memory
- Amount of available training data
- Target modality (or modalities)

In this section we characterise the performance of a gradient orientation detector for different feature sets. The sagittal MRI-T1 training cohort to used to train the detectors, and the (hitherto unseen) sagittal MRI-T1 *test* cohort is used for validation.

Method

Three detectors are demonstrated, where the goal is accuracy, generalisability (to different modalities) and memory usage respectively.

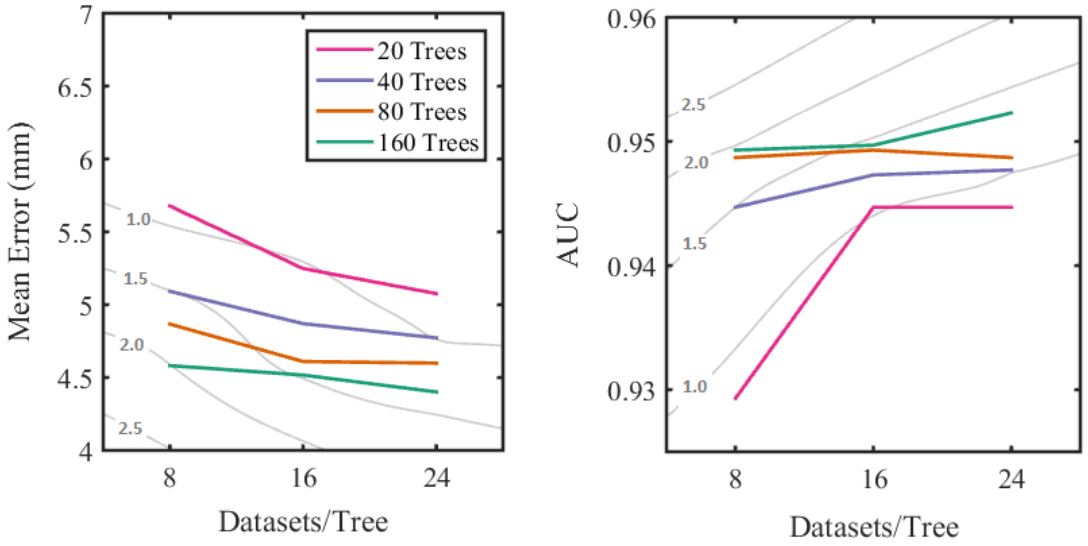
- Maximising Accuracy: Firstly, a detector is trained with the optimal parameters as found in section 3.4.
- Maximising Generalisability: Secondly, a detector is trained on *unsigned* gradients (gradients measured over a 180 degree range rather than over a 360 degree range) using half the number of bins, $B = 4$. This detector should generalise to different image modalities, and to contrasted and non-contrasted scans.
- Minimising Memory Usage: Finally, a detector is trained using unsigned gradients in the axial plane only ($B = 4$, $\psi = axial$). Such a detector might be useful where memory usage is a concern. We choose axial features since this has been shown to give best performance.

Reporting of results

Each experiment is run three times with three different random seeds, and the *mean* values are used to plot graphs showing the number of trees and the number of datasets per tree against accuracy.

The trade-off with time is illustrated by the grey contours on the graphs, which indicate detection times from 1.0 through to 2.5 seconds, at intervals of 0.5 seconds. These lines have been drawn by using a thin plate spline function to interpolate between the detection times for the 12 data points. This is an approximation, since the number of data points is small and the shape of the spline function takes no account of the nature of the (likely nonlinear) dependencies of detection time on T and on D_T. Accuracy could be increased with the addition of more data points. However, an approximation is sufficient for working purposes.

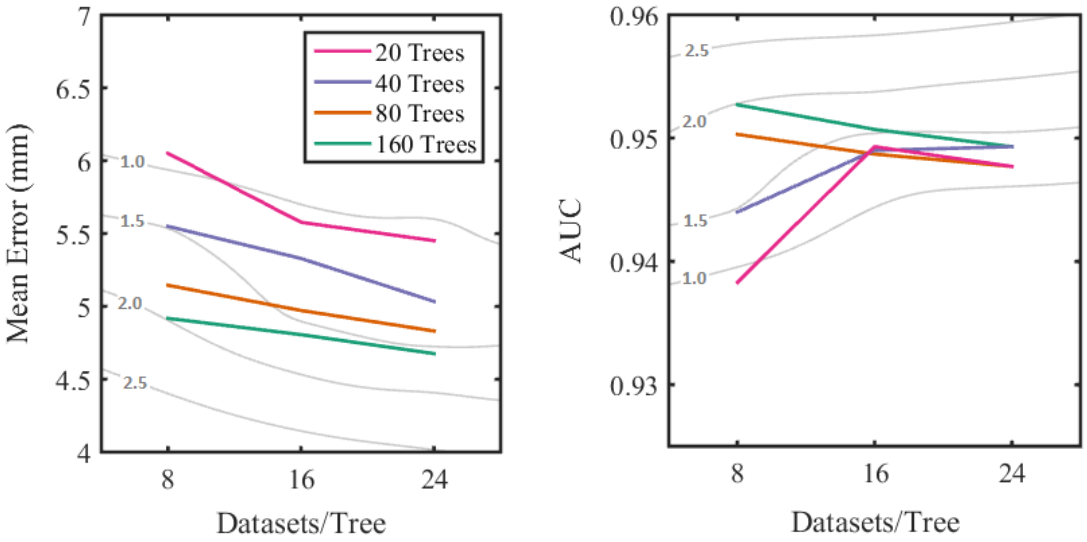
Given constraints of accuracy, run time, or even training time (more trees take longer to train, trees with more datasets take longer to train), such a graph could be used to select a detector with the optimum numbers of trees and datasets per tree.



(a) Mean Error

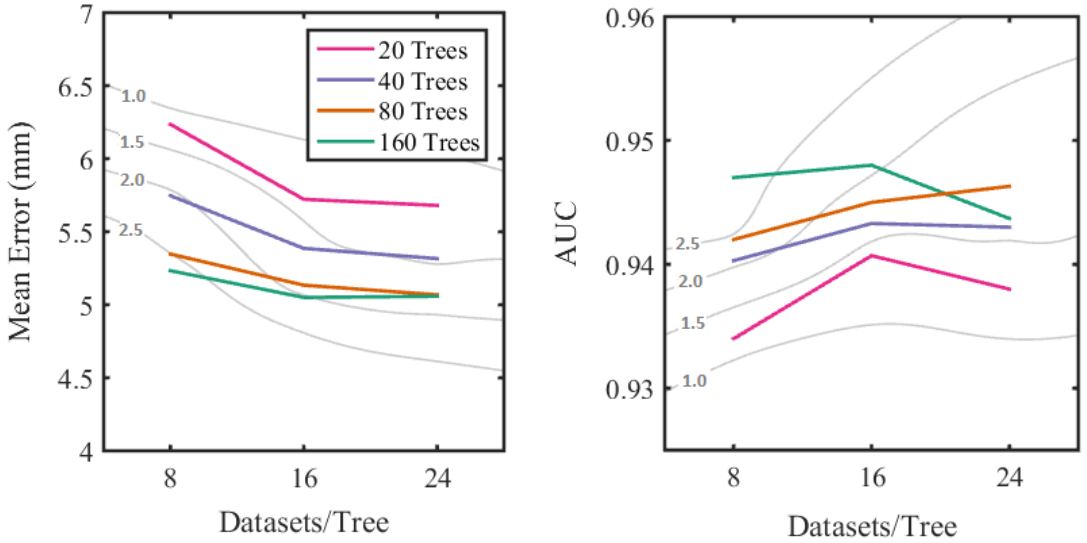
(b) AUC

Figure 3.22: Graphs showing datasets per tree D_T and forest size T versus accuracy for a signed gradient orientation detector. The trade-off with time is illustrated by the grey contours on the graphs, which indicate detection times from 1.0 through to 2.5 seconds, at intervals of 0.5 seconds.



(a) Mean Error (b) AUC

Figure 3.23: Graphs showing datasets per tree D_T and forest size T versus accuracy for an unsigned gradient orientation detector. The trade-off with time is illustrated by the grey contours on the graphs, which indicate detection times from 1.0 through to 2.5 seconds, at intervals of 0.5 seconds.



(a) Mean Error (b) AUC

Figure 3.24: Graphs showing datasets per tree D_T and forest size T versus accuracy for an unsigned gradient orientation detector, using features in the axial plane only. The trade-off with time is illustrated by the grey contours on the graphs, which indicate detection times from 1.0 through to 2.5 seconds, at intervals of 0.5 seconds.

Discussion

There are a number of effects seen on the graphs, which are discussed under the headings of accuracy and time. The relationship between T, D, D_T and T_{Unseen} was given in equation 2.23.

Accuracy

As T increases, accuracy also increases, although this asymptotically approaches a limit.

As D_T is increased, the performance of individual trees improves, since a greater amount of data is seen by each tree. However, the degree of randomisation across the forest is decreased. Hence the rate of improvement slows, and performance may even decline, as $D_T \rightarrow D$.

There is an additional negative effect when atlas location features are employed. Increasing D_T causes a reduction in the effective forest size when running out-of-bag detection on the training data cohort, which as noted above, would reduce the accuracy of the training data results. For instance, if $D_T = 24$ is chosen (with replacement) from a cohort of $D = 50$, then a forest of $T = 20$ has an effective leave-one-out forest size $T_{Out-of-bag} = 12$. This impacts on iterations 1 onwards where forests are trained using atlas location feature values derived from the training data results, using atlas mapping thresholds τ_P and τ_E which are also learnt from the training data results.

Time

According to the basic theory of random forests, it would be expected that the detection time:

- Increase proportionally to T
- Increase as D_T (and thus tree size) increases.

As D_T is increased, the detection time (counter-intuitively) decreases due to the effect of the forest shortcut (see section 2.3.2.6). More data per tree leads to more accurate trees, which means that fewer trees are required to be evaluated before background voxels can be definitely classified as such. However, as $D_T \rightarrow D$, and the data given to each tree becomes more similar, the ensemble of trees loses some of the randomness which makes the forest detector so effective. Hence the value of each additional tree is reduced, and it may be required to evaluate more trees before background voxels can be rejected.

3.6 Gradient orientation features in whole-body CT

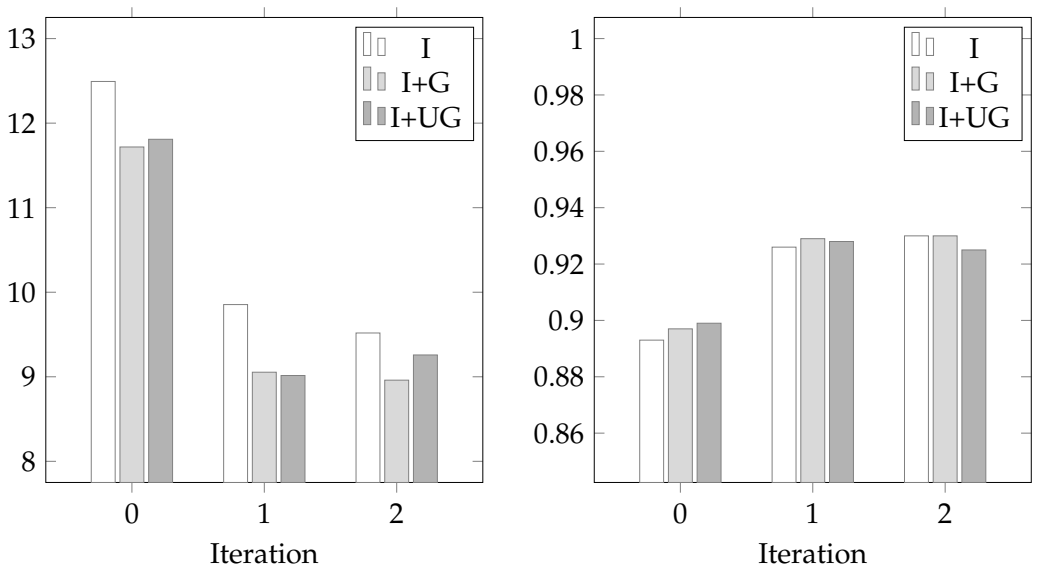
Description: We look to see if using a 50-50 mix of gradient orientation and intensity features aids in the whole-body CT landmark detector of chapter 2. Both signed and unsigned gradient orientation features are tried.

Method: As in chapter 2, we run the experiment for two iterations, using T = 80, D = 272, D_T = 40, F_T = 2500, radial sampling, d_{max} = 52mm, B = 8 (or B = 4 for unsigned), C_{max} = 30mm and ψ = {*axial, coronal, sagittal*} (one plane per tree).

Each experiment is run three times and the mean metrics are reported.

Results: The results are shown in Figure 3.25. Gradient orientation features do make a small but significant difference to the mean error compared to intensity features alone, of approximately 0.8mm. The difference between using signed and unsigned features is negligible.

Discussion: The improvement in CT is modest. This tallies with the earlier experiment in section 3.4.7.



(a) Mean Error (b) AUC

Figure 3.25: Graphs comparing the accuracy of intensity only (I), mixed intensity + signed gradient (I+G), and mixed intensity + unsigned gradient orientation features (I+UG), for landmark detection in whole-body CT data. The horizontal axis indicates iterations {0,1,2} i.e. two iterations of atlas coordinate feedback.

3.7 Mixing modalities: Cross-validation in MRI-T1, MRI-T2 and CT head scans

Description: The idea of mixing data of different modalities is of interest for a couple of reasons. Firstly, a one-size-fits-all general detector can be trained that should theoretically work for any modality, including previously unseen modalities. This is especially useful for MRI, where there are many different sequences. Secondly, where there is a lack of training data, training data from different sequences may be pooled. At TMVS, we have few training datasets for MRI sequences other than T1 and T2.

In this experiment, we look at CT, MRI-T1 and MRI-T2 modalities. T1 and T2 sequences differ in the depiction of fat and water. Fat (e.g. the myelin in white matter myelinated axons) is bright in T1 and dark in T2. Water (e.g. cerebrospinal fluid compartments, oedema) is dark in T1 and bright in T2. The two sequences are complementary, and the choice of scan depends on the anatomy or pathology of interest.

Detectors are trained with *unsigned* gradient features, and thus should in theory be blind to inversion of the image intensities (as was highlighted in Table 3.1), with the caveat that this does not hold at three-tissue boundaries (see Figure 3.4). We train three detectors on one modality each of MRI-T1, MRI-T2 and CT. The detectors are then cross-validated on each cohort to evaluate the inter-modality accuracy versus the intra-modality classification accuracy.

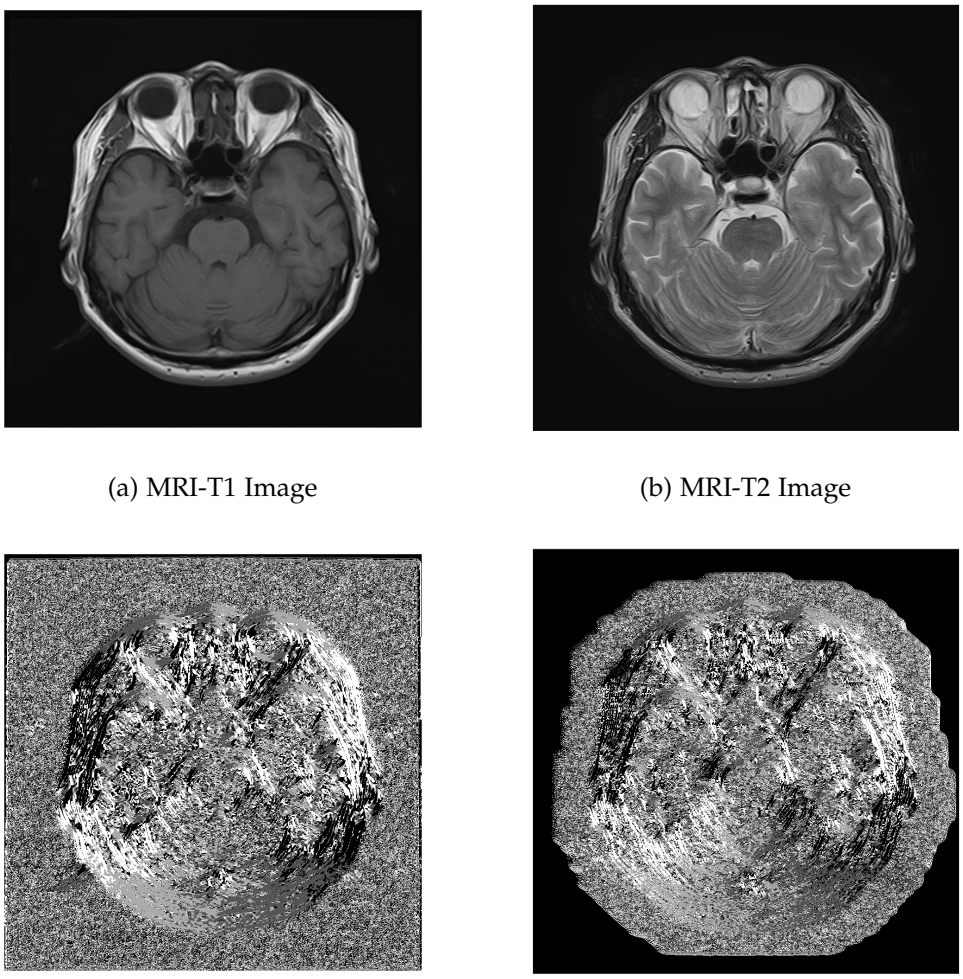
Method: There are different numbers of datasets in each modality cohort. To avoid the volume of training data being a confounding factor, a subset of 35 datasets are randomly selected from each cohort for use in this experiment i.e. the same number as the size of the smallest cohort, which is the axially acquired MRI-T1 cohort (see Figures 3.7 and 3.6).

For each random forest detector, we use $T = 160$, $D_T = 8$, $F_T = 2500$, radial feature selection, $C_{max} = 30$mm, $d_{max} = 52$mm and $\psi = \{axial, coronal, sagittal\}$ (one plane per tree). Unsigned gradient features ($B = 4$) are used.

Each experiment is run three times and the mean metrics are reported.

Result: The results are shown in Figure 3.27. Cross-modality classification appears to be eminently feasible, although some degradation is seen relative to intra-modality classification.

Discussion: The results are confounded by the fact that the acquisition regions of the different cohorts which we have at our disposal are different. In general, the CT scans are inclusive of the whole head and neck area. However, the MR datasets tend to focus on either the brain, the head, or the head/lower neck (with far fewer examples of the latter; there is only one example in the MRI-T1 group and none in the MRI-T2 group which show the neck cervical vertebrae). Thus if anything, the cross-modality results may appear worse than the true figures.



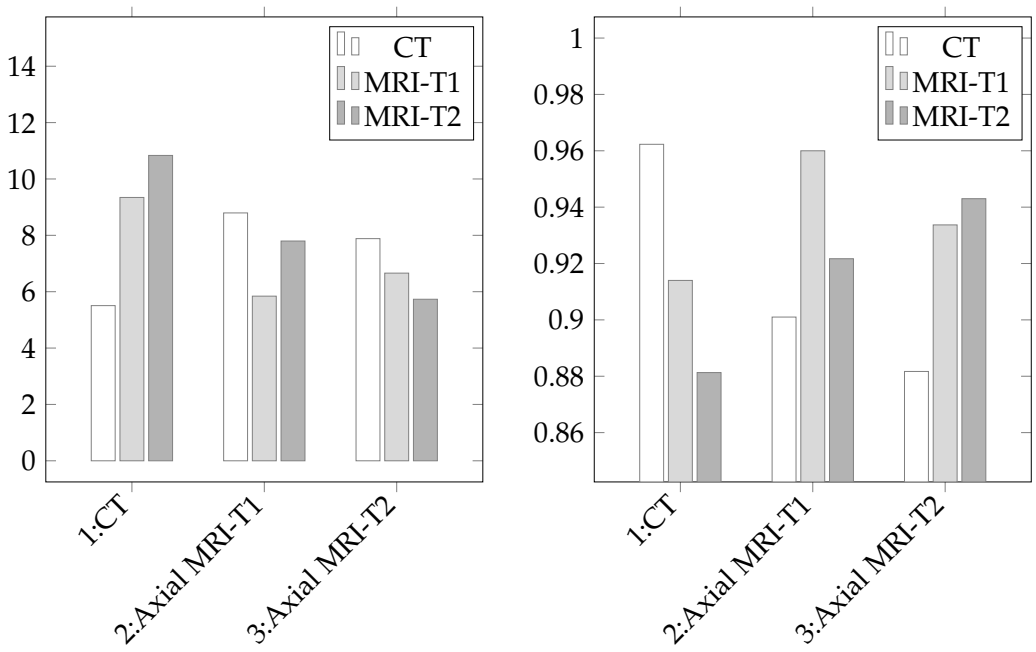
(a) MRI-T1 Image

(b) MRI-T2 Image

(c) MRI-T1 Unsigned Gradients

(d) MRI-T2 Unsigned Gradients

Figure 3.26: Images comparing the unsigned gradient orientations for equivalent MRI-T1 and MRI-T2 slices i.e. from the same patient acquisition. It can be seen that some of the colours are inverted in the original slices, but the gradient orientation representations appear very similar. [TMVS Dataset ID: 3538]



(a) Mean Error (b) AUC

Figure 3.27: Graphs showing the cross-validation of three detectors trained on three different modalities, using unsigned gradient features. The horizontal axis of the graphs is labelled with detector 1-3 and the modality on which they were trained. The bars above each represent that detector's landmarking performance on the three cohorts. It is evident that each detector performs best for the modality on which it was trained.

3.8 Discussion

3.8.1 Summary of our contribution

This chapter has shown that gradient orientation features have great promise when working in different MRI series images and in CT images. The ability to train a detector on one modality and then apply to a different modality is especially useful. Firstly, there is no need to train lots of specialised detectors for all modalities: rather one robust, general detector can be trained. Secondly, limited quantities of ground truth data from different modalities can be pooled. Thirdly, a detector can be trained which can work on images from previously unseen modalities.

The refinements which are present in established gradient-based descriptors such as HOG and SIFT were found to have no significant benefit, such as weighting schemes in spatial or orientation space, and trilinear interpolation in orientation space. It was also found that a small number of bins gave close to optimal performance. Rotational invariance (an integral part of the SIFT transform) was not attempted since the orientation of the patient in scanner space has meaning. In many aspects, simplicity has won over complexity in feature design.

3.8.2 The relationship between total feature space size and feature subspace size

It is interesting that we found the optimum number of features per tree F_T was about the same for gradient orientation features as for the single-voxel intensity features. The size of the gradient orientation feature space is about two orders of magnitude greater than the number of intensity features (see Table 3.2, details of the calculations are given below). Consequently, we might expect the optimum value of F_T to be proportionately much larger. However, there is redundancy in the feature space:

- There is overlap in the set of cuboids C over which the features are computed.
- Since the bin probabilities sum to 1.0, there is actually one less degree of freedom than B, so the *effective* feature space size is $F \times \frac{B-1}{B}$.

Feature	F
$I(v+d)$	15625
$I(v+d) - I(v)$	15624
$f_{orient}(b, v, d, C, \psi)$, 8 bins	11,998,550
$f_{orient}(b, v, d, C, \psi)$, 4 bins	5,999,275

Table 3.2: Comparison of the sizes of the feature spaces for intensity features and for intensity gradient orientation features. Numbers are computed computed at D_{Res} = 4mm voxel^{-1}, see text for computations. The *relative* intensity at an offset of $d = (0, 0, 0)$ is always zero, so this is a redundant feature, hence we show the feature space as being one less than for absolute intensity features. We use 4 bins for unsigned gradients, so the feature space halves in size.

We suggest that this redundancy accounts for the vastly reduced proportion of the feature space that we require to sample per tree, compared to intensity features.

The feature space computations are laid out below.

Intensity features

$$\begin{aligned} F &= \left(2 \times \frac{d_{max}}{D_{Res}} - 1\right)^3 \\ &= \left(2 \times \frac{52}{4} - 1\right)^3 \\ &= 15625 \end{aligned} \tag{3.2}$$

Gradient orientation features

$$\begin{aligned} F &= B \times (\frac{C_{max}}{D_{Res}})^3 \times \frac{4}{3}\pi \left(\frac{d_{max}}{D_{Res}}\right)^3 \\ &= 8 \times (\frac{32}{4})^3 \times \frac{4}{3}\pi \left(\frac{52}{4}\right)^3 \\ &= 11,998,550 \end{aligned} \tag{3.3}$$

3.8.3 Atlas location features versus gradient orientation features

It is interesting — and heartening — that gradient orientation features have given additional benefit on top of the atlas location features introduced in chapter 2 for autocontext.

To further explore this result, we extend the experiment from section 2.5.2, running a landmark detection experiment using the *ground truth* landmarks for the mapping (only one iteration required). The results are as shown in Table 3.3. We include radial intensity features since these this pattern of sampling is used in the gradient orientation experiments; there appears to be no significant difference between radial and volumetric sampling in the case of intensity features.

Mapping	**Features**	**Mean Error (mm)**	**AUC**
Affine	$T_a(v)$	12.71	0.890
Affine	$T_a(v)$, $d_{sag}(v)$	12.66	0.897
Affine	$T_a(v)$, $d_{sag}(v)$, $I(v+d)$	8.17	0.955
Affine	$T_a(v)$, $d_{sag}(v)$, radial $I(v+d)$	8.12	0.950
Affine	$T_a(v)$, $d_{sag}(v)$, radial $I(v+d)$, $f_{orient}(b,v,d,C,\psi)$	7.62	0.951
Spline	$T_a(v)$	2.36	0.976
Spline	$T_a(v)$, $d_{sag}(v)$	2.43	0.980
Spline	$T_a(v)$, $d_{sag}(v)$, $I(v+d)$	3.64	0.990
Spline	$T_a(v)$, $d_{sag}(v)$, radial $I(v+d)$	3.90	0.990
Spline	$T_a(v)$, $d_{sag}(v)$, radial $I(v+d)$, $f_{orient}(b,v,d,C,\psi)$	3.92	0.990

Table 3.3: Results of landmark detection using mappings created from the ground truth. This is an extension of Table 2.5. Spline = Thin plate spline (similarity). Results are the means of three experiment runs, run with different randomisation seeds.

It can be seen that when the mapping is perfect, as is the case with the spline mapping, gradient orientation features provide no extra information. On the other hand, in the case of the affine transformation there is a small improvement. This supports the hypothesis that gradient orientation features are providing a small amount of useful information in CT data, over and above that which can be provided by either an affine mapping or simple intensity features.

3.9 Future work

3.9.1 Pooling data using unsigned orientations

An interesting question is to what extent does the addition of training sets from modality A improve results on modality B? Would results improve even where plenty of training data is available for modality B, perhaps by encouraging the choice of robust features? In concrete terms, the experiment to be conducted would be to plot in two dimensions the accuracy of a system applied to modality A, for a range of training set sizes from A and B.

At the heart of machine learning algorithm design is the *bias-variance tradeoff*. Algorithms exhibiting high *bias* underfit the training data, and do not capture all of the salient relationships in the data. This manifests as poor accuracy in both training and test data. Algorithms with high *variance* overfit to the training data, and do not generalise well to unseen data. This manifests as much better accuracy in the training data, compared to the test data, or, a large change in the algorithm in response to a small change in the training data. Random forests have lower variance compared to the variance of a single decision tree, but also have slightly higher bias. Usually the former outweighs the latter, giving a better overall trade-off.

Greater amounts of training data lead to further reduced variance. Where data is scarce, the pooling of data from different modalities will lead to a larger and richer training set which yields a correspondingly improved random forest. The question of whether accuracy would improve when data is already plentiful for any one modality depends on whether the algorithm has too high variance or too high bias. In the case of excessive variance, a different modality may encourage the selection of more robust features which are universal in both modalities and reflect real anatomical detail. In the case of excessive bias, or optimal variance-bias trade-off, the inclusion of data from a different modality runs the risk of discouraging the selection of features which have meaning in modality B but not in modality A.

This is of practical relevance to us at TMVS, since it has potential to improve accuracy in MRI by adding in our wealth of CT training sets, and is a priority for future experimentation.

Chapter 4

Probabilistic fusion for automated segmentation of brain structures

Abstract

This chapter adapts the random forest described in 2.3.2 for the purpose of brain gyrus segmentation. We try three variant random forest classifiers, using contrasting sets of the intensity and gradient orientation features described in chapters 2 and 3. In isolation, these classifiers are shown to be far inferior to a multi-atlas segmentation (MAS) method. However, we are interested in seeing if it is possible to combine the MAS and forest classifiers, which appear to have complementary strengths and weaknesses, to make a superior hybrid classifier. To this end, we show segmentation results from two simple, non-parametric, classifier combination operations: simple averaging and the Bayesian product. A final empirical calibration step is employed to calibrate the predicted class probability values according to the observed frequencies in the training data. We show that, when calibration is used, the hybrid classifier gives a marginal quantitative improvement for the datasets in the MICCAI 2012 brain segmentation challenge compared to the MAS classifier alone, and further show qualitatively that this is due to the forest giving sharper delineation between grey matter and white matter at the boundaries. Whilst our hybrid classifier is not the best performer compared to other authors who have published on this dataset, it has far greater speed than the top methods, and thus demonstrates good compromise between accuracy and speed. To increase confidence in the significance of the small Dice Score improvement, two further relevant metrics are computed which also show improvement, and by also comparing with contemporary results on this dataset we conclude that we are nearing the limits of feasible accuracy. In summary, classifier combination is a promising area, however the need for empirical calibration in our experiments shows that more complex combination methods than averaging or Bayesian multiplication merit investigation.

4.1 Synopsis

In this chapter we

- (4.3.3) Adapt the random classification forest used for anatomical landmark detection for the purpose of brain structure segmentation.
- (4.3.4) Present two simple methods (averaging and Bayesian product) of combining the forest classifier with a pre-existing multi-atlas registration classifier, to give improved hybrid classifiers.
- (4.3.5) Introduce a calibration step, to transform classifier probability values into true (empirically observed) probability values.
- (4.4) Benchmark all algorithms against the results of the MICCAI 2012 brain segmentation challenge.
- (4.4) Show that the optimal choice of random forest features differs depending on whether the classifier is used alone or in combination with an atlas classifier.
- (4.4.4) Discuss the effect of handling prior class probabilities at the tree level (as done for landmark detection) or at the forest level, as we do in this chapter.
- (4.4.5) Increase confidence in the significance of the Dice Score results by computing two further relevant metrics and demonstrating positive correlations between the metrics.
- (4.5.2) Discuss the importance of independence when combining evidence.

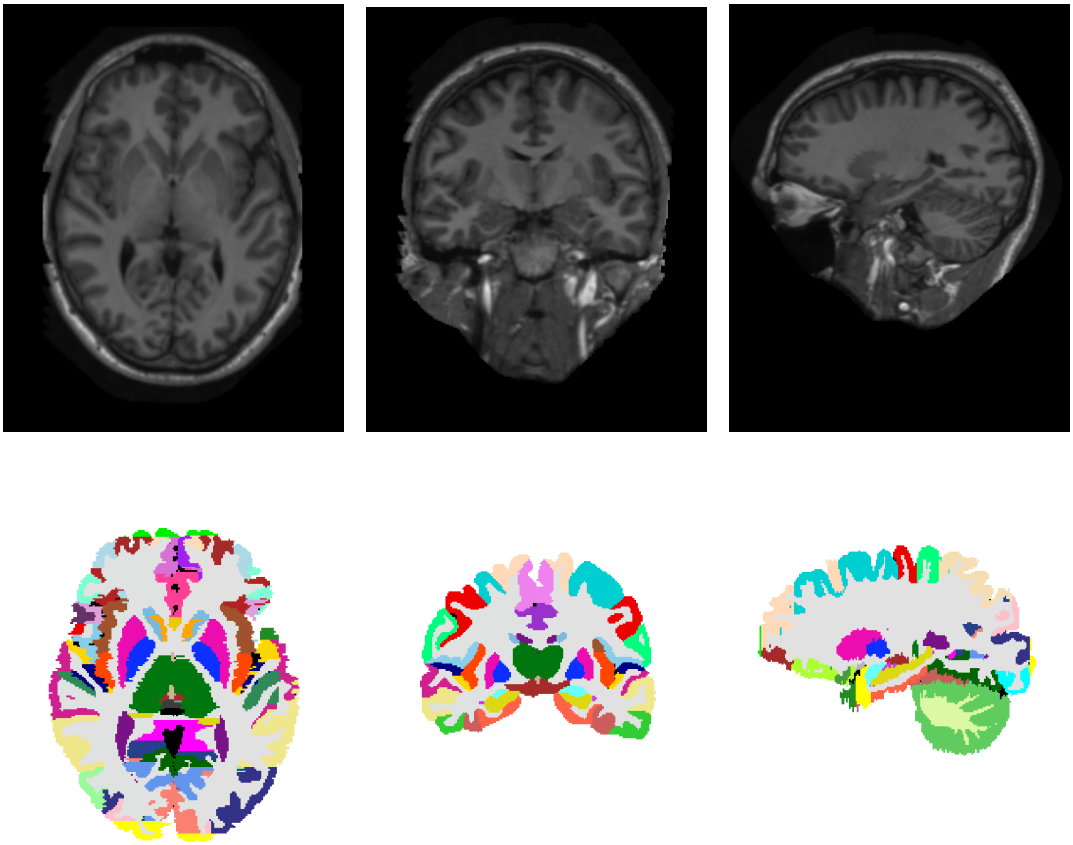


Figure 4.1: Axial, coronal and sagittal mid-volume slices from an example brain dataset. The T1-sequence MRI data is shown on the top row, and the corresponding ground truth is shown below. The brain has been fully labelled from the 138 regions defined in the Neuromorphometrics protocol. Each colour corresponds to a different label. [TMVS Dataset ID: 9901]

4.2 Introduction

4.2.1 Problem definition

Identification and segmentation of brain structures plays a role in visualisation of brain anatomy, localisation of pathology, volumetric measurement of specific brain structures, and planning of surgical or radiotherapy procedures. In this chapter we tackle the task of automatically segmenting the brain into 138 functionally distinct regions as per the Neuromorphometrics protocol[99].

An example dataset is given in Figure 4.1.

4.2.2 Prior art

Parcellation of the brain is a difficult problem, even for manual observers. There are 138 regions and only three distinct tissues types — grey matter, white matter and cerebrospinal fluid (CSF) — so many regions are made of the same tissue type. Boundaries between regions of different tissue types (in particular between grey matter and white matter) are often intricate and indistinct. Boundaries between regions of the same tissue type are often invisible; there is sometimes an obvious point of separation (e.g. partitioning of the grey matter at the sulci), but in many cases boundaries have been arbitrarily designated as the straight line connecting two observable landmark points. Finally, one gyrus may look very like another. Therefore, spatial context is key.

Segmentation techniques based on intensity and shape information such as morphology, deformable surfaces, level sets, watershed and graph cuts are not a perfect fit as they are not sensitive to invisible region boundaries. Coarse segmentation of the three tissues in itself is problematic since the intensity distributions overlap [100] and there are additional phenomena to contend with such as noise, partial volume effects and bias field artefacts.

Atlas-based techniques have therefore gained in popularity for this kind of problem. Given a reference image manually labelled with the structures of interest (i.e. an atlas), segmentation may be treated as a registration problem from the atlas volume to the novel volume, using one of the myriad registration techniques that exist, thereafter simply transferring the labels according to the discovered mapping [101, 102, 103, 104, 105].

The single atlas approach has the disadvantage that the atlas may not correspond particularly well to the novel image. Combining multiple atlases in some way should better model anatomical variation. Probabilistic atlases [106] were the first attempt to do this. Multiple atlas volumes are registered into a standard coordinate space, and voxelwise population intensity mean and variance statistics can then be computed. *Voxel-based morphometry* [107] is an analysis technique from the 90s whereby many brains are registered to a template, and statistical parametric mapping is used to make voxelwise comparisons. Fischl [100] presented a Bayesian framework for subsequent classification into anatomical regions. The prior for a class c was computed based on the proportion of atlases mapping a voxel labelled as c to the given point in atlas space. The intensities were modelled as Gaussian distributions, one per class. The *maximum a posteriori* class was then chosen at each voxel based on the class prior and the intensity

likelihood given the class. Van Leemput [108] demonstrated an expectation maximisation approach where the novel dataset was first *affinely* registered to the space of a probabilistic atlas, giving initial classification probabilities. The intensity parameters and voxel classifications were then estimated iteratively, along with parameter estimation for the MRI bias field correction and a Markov Random Field representing spatial constraints.

Fischl [67] and Ashburner and Friston [43] went on to develop the idea of coupling the deformable registration step (to atlas space) with the estimation of the intensity distributions, using a common objective function to jointly estimate deformation and intensity parameters. When the atlases are warped according to the discovered deformation parameters, this provides voxelwise spatial prior as to the class. The tissue intensity parameters are computed as per-class Gaussian distributions. Fischl computed these per class and per atlas location, whereas Ashburner and Friston computed these as stationary distributions for grey matter, white matter and cerebrospinal fluid (but with more than one Gaussian allowed per tissue, resulting in more than three classes). Both models also included bias field parameters, and as mentioned in chapter 3 Fischl further explored the physics of MR imaging in order to confer invariance to acquisition parameters.

Multi-atlas segmentation (MAS) methods [109, 110, 111, 112] are slightly different in that each atlas is registered to the novel image, and the labels are then aggregated, either by a simple majority vote or by some more complex procedure to give a consensus result. In contrast to a probabilistic atlas, MAS methods allow the weighting of the atlases to be varied (or for selective use of atlases [113, 111]) such that those estimated to be more similar to the novel data are given greater weight. Since registration is performed directly to the novel image, rather than to some (perhaps arbitrarily chosen) reference image, or to some averaged intensity image, we also expect an optimal registration result. A number of MAS variations were showcased in the 2012 MICCAI Grand Challenge on multi-atlas labelling of the brain [1]. We refer the reader to Iglesias and Sabuncu [114] for a comprehensive review of MAS approaches up to the end of 2014. Sources of inaccuracy include the (imperfect) quality of the registration, and the voting bias towards the mean where a straightforward majority voting rule is used. Hence the development of local fusion schemes [115, 116, 117, 118, 119, 120, 121] where atlas contributions are weighted on a local basis, and post-processing steps [112, 22, 122] to retrospectively correct for registration errors.

For instance, many label fusion schemes [117, 121, 118, 119, 120] are based on STAPLE (Simultaneous Truth And Performance Level Estimation) [123]. In

STAPLE, the quality of a set of segmentations — which may be ground truth acquisitions — is simultaneously evaluated, using expectation maximisation to iteratively find segmentation weighting parameters which maximise the data log likelihood. The data likelihood is expressed as a function of intensity, spatial distribution of structures, and the agreement between segmentations. By contrast, Wang *et al.* [22] advocated a machine learning post-processing step to correct for systematic biases. An Adaboost classifier was provided with a set of features including local neighbourhood intensities and estimated segmentation labels in addition to the spatial coordinate (relative to the estimated centre of mass of the region of interest). Van der Lijn [124] proposed a hybrid method of registering all atlases to the novel image in order to create a probabilistic atlas. In a post-processing step, graph cuts were then used to solve the energy functional formed from the spatial prior energy term yielded by the probabilistic atlas, and a regularisation energy term introduced to produce smooth segmentations. Other methods have attempted to explicitly model atlas errors and codependencies [125, 126].

The practical drawback for state of the art registration methods is the long run time, usually of the order of hours or even days [1], because MAS methods require as many registrations as there are atlases, and state-of-the-art deformable registration methods are slow. One option is to invest in more powerful hardware; see [127] for a review of medical image segmentation algorithms on GPUs. However, a number of approaches have been proposed as alternatives. Asman *et al.* [128] trained AdaBoost learners to mimic the result of MAS, mapping from an initial weak segmentation result to the equivalent of MAS segmentation. All atlases and the novel were initially registered affinely to a template atlas space. Then for the weak segmentation, atlases were selected for the particular dataset according to pairwise similarities in a low-dimensional data space (obtained by principle component analysis of the non-background voxels). Patch-based segmentation [129, 130, 131] also requires only affine registration in order to identify a constrained search window centred at each voxel, within which all similar patches from all atlases are considered for estimation of the true voxel label. Bai *et al.* [131] demonstrated improved performance by augmenting patch intensity information with gradient and contextual features. Wang *et al.* [132] bypassed registration completely for knee MRI image segmentation by starting with large patches at a coarse resolution and working down to finer resolutions. Wang and Yushkevich [133] bypassed registration for brain tumour segmentation by segmenting into supervoxels [134] and matching similar supervoxels to

find an initial labelling, before descending to voxelwise segmentation by patch correspondence.

A pure random forest approach is fallible for the same reason as other appearance-based methods; that we are considering many very similar structures which are difficult to differentiate on the basis of appearance alone (although a random forest has been demonstrated by Iglesias *et al.* for the extraction of the brain as a whole [135]). Any of the spatial context methods discussed for anatomical landmark detection could be explored, for instance entangled forests [15], structured random forests [52], autocontext [54, 28, 136] or indeed a learnt atlas coordinate feature. Additionally, in a global optimisation model for cortical sulci segmentation, Tu *et al.* [137] use dynamic programming to optimise an energy equation formulated from a shape prior and the output of a probabilistic boosting tree.

Other approaches combine multi-atlas and forest methods. Zikic *et al.* [21] registered the novel dataset to a reference, to which the atlases had been pre-registered, and subsequently ran classification using a set of *atlas forests*. One forest was trained on each atlas using the local neighbourhood atlas label priors as features alongside typical image features. This method was demonstrated on several brain segmentation problems. Gauriau *et al.* [23] combined probabilistic atlases (one per organ) with a regression forest. The forest was trained to estimate affine registration parameters for each organ. The probabilistic atlas was then registered multiple times, according to every voxel's estimate of the location and scale of the organ, and with a weighting corresponding to the voxel's confidence. The accumulated result was termed a *confidence map*, and used as the template for subsequent segmentation by template deformation.

4.2.3 Motivation for our approach

Given that we already have an established fast MAS registration algorithm at TMVS [26] and a mature random forest classifier algorithm as described in section 2.3.2, we propose to leverage the power of both by training two independent classifiers (one MAS classifier and one forest classifier), and combining the results. We hypothesise that this approach will exploit both spatial-based and feature-based information, to produce a robust hybrid classifier which gives greater overall accuracy. The hybrid classifier should still have good speed since both are fast algorithms.

Such a combined classifier could be termed an ensemble classifier, albeit

minimalist, given that the ensemble consists of just two classifiers. In fact, the two base classifiers are themselves ensemble classifiers, so there will be two levels of combination. Methods of classifier combination have been reviewed by Kittler *et al.* [138], Kuncheva [139] and Tulyakov *et al.* [140]; these are many and varied. We evaluate the results of adopting two alternative combination rules that are fast to compute: Bayesian product combination and simple averaging of the base classifier probabilities.

To clarify, Bayesian product combination effectively refers to the *product rule*, but we take account of class priors. Similarly, simple averaging is equivalent to the *sum rule*. Both are distinct from *Bayesian model averaging* [141] which refers to weighted averaging of classifier results, each being weighted by the posterior probability that the classifier itself is the true model according to the data. In theory, this should be computed over the space of all possible models (or classifiers). In practice, the model space is sampled. Either way, this approach is unsuitable when considering just two classifiers but we mention it to avoid confusion since it is a commonly used Bayesian approach.

It should be noted that simple averaging is uncontroversially employed *within* the base classifiers, when combining the atlases and trees respectively.

In a final step, we look to the philosophy proposed by Dawid [142] for calibration of Bayesian probabilities. Empirical calibration is applied independently for each class, to modify the hybrid classifier probabilities to match the real probabilities as measured empirically from the training data. Hence, in our evaluation, the simple combination rules are somewhat enhanced by this final empirical correction; both calibrated and uncalibrated results are given for comparison.

Experiments are also presented using three contrasting mixes of intensity and gradient orientation features for the random forest classifier, in order to investigate which features perform best. This furthers our work from section 3.4.7, in comparing and contrasting intensity and gradient orientation features.

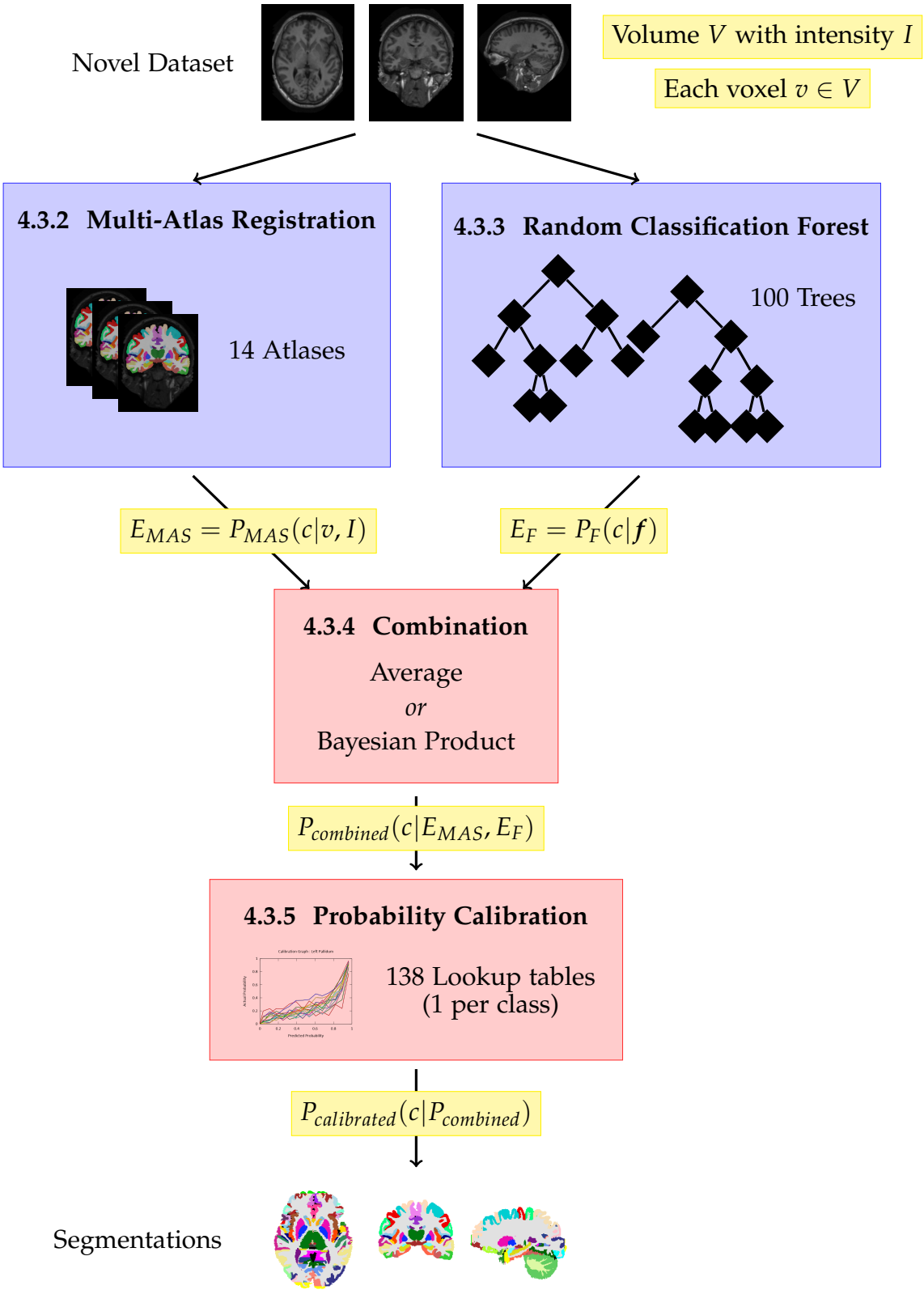


Figure 4.2: Overview of the brain segmentation method. The evidence E from the MAS and forest classifiers, the class probabilities P_{MAS} and P_F, are combined into a hybrid classifier. This chapter focuses on the *combination* and *calibration* components (marked in pink).

4.3 Method

Figure 4.2 shows an overview diagram of the brain segmentation method. In short:

- A multi-atlas registration algorithm and a random classification forest are trained independently.
- The two sets of posterior probabilities are combined using both the *average* and (on the assumption of classifier independence) the *Bayesian product*.
- The combined probabilities are calibrated using transfer functions which are empirically determined from the training data.
- Finally, segmentation labels are generated by selection of the class with maximum probability at each voxel.

More detail about each stage follows.

4.3.1 Data and ground truth collection

We used the data provided for the MICCAI 2012 Grand Challenge [1] in order to benchmark against other algorithms. See Figure 4.1 for an example. This data consists of de-faced T1 MRI brain volumes from the Oasis database, which are pre-processed as follows.

- Each original volume is an average created by co-registering two scans and resampling to 1 mm isotropic voxels.
- Automated bias field inhomogeneity correction is applied.
- The volume is re-oriented and aligned using three manually placed landmarks.
- The volume is resliced according to the new alignment.
- Brain extent landmarks are chosen and the scans are cropped to make efficient use of computer RAM and screen area.

The training data comprises 15 scans from 5 males and 10 females, with a mean age of 23yrs (range 19 to 34 yrs). The test data comprises 20 scans from 15 distinct subjects (with 5 repeats — no details available about the reasoning behind the

repeats). The subject group consists of 10 females and 5 males, with a mean age of 38yrs (range 18 to 80 yrs). Some scans have signs of pathology.

Ground truth comprises manual segmentations of 138 brain structures, created according to the Neuromorphometrics protocol [99].

4.3.2 Multi-atlas registration

The first classifier is a fast multi-atlas registration algorithm by Murphy *et al.* [26]. Each atlas dataset is registered to the novel dataset using affine registration followed by Demons-based deformable registration [143]. The similarity metric is mutual information [144, 145]. The atlas registration probability, $P_{MAS}(c|v, I)$, that c is the class given the position of the voxel of interest v and the intensity data I in the novel dataset, is the mean vote over the A atlases.

$$P_{MAS}(c|v, I) = \frac{\sum_{i=1}^{A} a_i(c|G_i(v|I))}{A} \tag{4.1}$$

$G_i(v|I)$ is the mapping from a voxel v in the novel dataset to its corresponding position in the atlas a_i, the mapping being dependent on the intensity information I in the novel dataset. The atlas returns a binary vote $a_i(c|G_i(v|I)) \in \{0, 1\}$ indicating if class c is the label at $G_i(v)$.

We do not use the expectation maximisation post-processing step described in [26] because this makes the simplistic assumption that each class can be modelled by a single Gaussian distribution. This does not hold in reality, since some classes contain mixtures of different tissue types and there are partial volume effects to take into account.

4.3.3 Random classification forest

The second classifier is a random classification forest based on that used for anatomical landmark detection (see section 2.3.2). Table 4.1 shows the forest parameter values which can be directly compared to the settings for the landmark detection classifier (see Table 2.1). The differences are described below.

4.3.3.1 Data sampling strategy

Experimentation with the training data cohort was used to determine a pattern of data sampling as follows.

Parameter	Definition	Value
D_{Res}	Resolution at which classifier is run	1mm voxel^{-1}
T	Number of trees in forest	100
D	Number of training datasets	15
D_T	Number of training datasets sampled per tree (bagging)	5
d_{max}	Maximum feature offset	Variable, see 4.3.3.2
F	Total number of possible features	Variable, see 4.3.3.2
F_T	Number of features selected per tree	Variable, see 4.3.3.2
$\sigma_{Sampling}$	Standard deviation of Gaussian weighting function for landmark samples	N/A
B_{Ratio}	Ratio of background to foreground training samples	0.1
B_{π_Ratio}	Ratio of background class to brain region class prior probability	Implicit, see 4.3.3.3
w_{Node_min}	Minimum total weight of samples in a node for branch splitting, otherwise branch is terminated.	10.0
$w_{Node\_Split\_min}$	Minimum total weight of samples in smallest child node, otherwise branch is terminated.	5.0
d_{skip}	(Detection phase) Grid search interval	N/A
T_{min}	(Detection phase) Minimum number of trees to evaluate before forest shortcut may be deployed.	N/A
$P_{Shortcut}$	(Detection phase) Minimum probability for forest shortcut.	N/A

Table 4.1: Parameter values for the random classification forest for brain segmentation. This table can be directly compared with Table 2.1 to see differences from the anatomical landmark detection classifier.

From each training dataset, ten thousand samples are selected randomly and uniformly across the brain region. Additional samples are subsequently chosen to boost small volume classes up to a minimum of 200 samples per class. Background samples are selected from the remainder of the dataset in a ratio of 10:1 foreground (brain) samples to background (non-brain) samples i.e. $B_{ratio} = 0.1$.

All samples have equal weighting of 1.0.

4.3.3.2 Mix of features

Since the MR image values are uncalibrated, we first linearly normalise volume intensities such that the 5th and 95th voxel intensity percentiles map to 0 and 1000 respectively (as in chapter 3).

Drawing on the results of landmark detection in MRI images, three types of features are trialled: relative intensity features, cuboidal relative intensity features and gradient orientation features.

- $I(v + d) - I(v)$: Single voxel relative intensity features, comprising the difference between the intensity of the voxel v itself and a voxel at a displacement d from v. The displacement d is a randomised parameter.
- $I(v + d, C) - I(v)$: Cuboidal intensity features, comprising the difference between the intensity of v and the mean intensity of a cuboid C centred at a displacement d from v. The displacement d and the size of the cuboid C are randomised parameters.
- $f_{orient}(b, v, d, C, \psi)$: Gradient orientation probability (according to a normalised 8-bin histogram of oriented gradients) within a cuboid centred at a displacement d from v. The orientation bin b, the displacement d, the cuboid size C and the anatomical measurement plane ψ are randomised parameters.

Displacements d are sampled uniformly with respect to magnitude, thus yielding a greater spatial density of features local to v (radial sampling). d is randomly selected between 1mm (which is the dataset resolution, and also the resolution at which the algorithm is run) and d_{max}. Each component (x, y, z) of C is randomly selected between 1mm and C_{max}.

Following experimentation on the training data, we present results for classifiers with three different combinations of features (classifiers A, B and C).

Decision Forest Classifier variants

Classifier A : F_T = 200. Single voxel intensity features. d_{max} = 15mm.
Classifier B : F_T = 1000. 50% of trees trained on long-range single voxel and 50% of trees trained on cuboidal intensity features. C_{max} = 30mm and d_{max} = 50mm.
Classifier C : F_T = 1000. 50% of trees trained on single voxel intensity features and 50% of trees trained on gradient orientation probability features. C_{max} = 30mm, d_{max} = 50mm, B = 8 and ψ = {*axial, coronal, sagittal*} (one plane per tree).

Fewer features are used for classifier A since the pool of features is smaller with d_{max} = 15mm than with d_{max} = 50mm (14k versus 524k, see computation in section 3.8.2).

4.3.3.3 Computation of leaf node probability distributions

For the brain segmentation classifier, a different approach is taken for computing the leaf node posterior probabilities compared to that for the landmark detection classifiers. This yields the true posterior probabilities without the need for re-weighting or prior estimation. This also reduces potential overfitting.

We pass ***all*** voxels from the D_T training datasets through the tree (not just those used to train the tree). For a novel voxel v, whose feature values are computed from the image data I, the tree probability $P_T(c|v, I)$ of class c is shown in equation 4.2. In words, it is the frequency $f_{Lx}(c)$ of training voxels belonging to c in the end leaf that v reached, divided by the total frequency of training samples in the leaf over all M classes.

$$P_T(c|\boldsymbol{f_t}) = \frac{f_{Lv}(c)}{\sum_{i=1}^{M} f_{Lv}(c_i)} \tag{4.2}$$

The forest probability $P_F(c|v, I)$ is computed by taking the mean of the distributions in the leaves of all T trees.

$$P_F(c|\boldsymbol{f}) = \frac{\sum_{t=1}^{T} P_t(c|\boldsymbol{f_t})}{T} \tag{4.3}$$

4.3.3.4 Treatment of missing feature values

Missing values occur when feature information is located outside of the scanned volume. In the case of cuboidal features, if more than half of the box lies inside of the volume then we compute the feature over the partially visible portion. For landmark classification (see section 2.3.2.3), samples with missing feature values were sent down both branches with a half-weighting. However, the Neuromorphometrics datasets are cropped quite closely to the brain extent. As a result there are many samples with many missing values, leading to a long detection time due to the large numbers of branches which must be traversed.

Hence we employ a strategy where, during tree traversal (both during training of the tree and during detection), samples with missing values are sent down the left hand branch of the tree. This strategy corresponds to assuming that the missing value falls below the feature threshold, which would usually be the case if the unknown voxels within the box represent air. It is a fast and simple method that is accurate when the position of the padding relative to the brain region is fairly consistent between volumes *or* when the assumption of air is reasonable, and is a good fit for the Neuromorphometrics datasets which are cropped using manually placed landmarks. This gives a run time of the order of 100 times shorter.

4.3.4 Classifier combination

Where classifiers have similar performance, then combining them will yield a more robust result, assuming that the errors are independent. If one classifier is more accurate for all classes than the other, then combining the two will *potentially* just dilute the expert result. However, this is not necessarily the case. For instance, one classifier may tend to overestimate the probability of a given class whilst the other underestimates the same class; in this case the combined estimate may be closer to the true probability than either individual classifier. We investigate to see whether there is value in combining the MAS and forest classifiers.

Our approach could be considered a form of ensemble classifier — an ensemble of just two. Ensembles are typically combined by probability averaging, due to the large degree of statistical dependence. We present the averaged result first.

Since our ensemble is of different classifiers, there are grounds for making the assumption that the classifiers are conditionally independent given the class. An alternative is then to multiply the results according to Bayes' theorem. Thus we secondly present the results from taking the product of the two classifiers, with

due consideration of priors.

Each of these methods is now defined, using the notation E_{MAS} and E_F for the evidence given by multi-atlas registration and the random classification forest.

Averaging

The average $P_{Av}(c|E_{MAS}, E_F)$ is the simple mean.

$$P_{Av}(c|E_{MAS}, E_F) = \frac{P(c|E_{MAS}) + P(c|E_F)}{2} = \frac{P_{MAS}(c|v, I) + P_F(c|\boldsymbol{f})}{2} \tag{4.4}$$

Bayesian product

The Bayesian Product posterior probability $P_{BP}(c|E_{MAS}, E_F)$ is proportional to the product of the two classifiers, divided by the class prior. A derivation follows, based on that presented in [146]. First, following Bayes' rule (we denote the prior by π):

$$P_{BP}(c|E_{MAS}, E_F) = \frac{P(E_{MAS}, E_F|c)P(c)}{P(E_{MAS}, E_F)} = \frac{P(E_{MAS}, E_F|c)\pi(c)}{P(E_{MAS}, E_F)} \tag{4.5}$$

We then make the assumption that the classifier results are independent of one another *given the class*. This assumption is justified due to the difference in their approaches: the decision forest is dominated by local intensity information whereas the atlas registration is dominated by spatial context. This leads to equation 4.6.

$$P_{BP}(c|E_{MAS}, E_F) = \frac{P(E_{MAS}|c)P(E_F|c)\pi(c)}{P(E_{MAS}, E_F)} \tag{4.6}$$

Finally, by transforming each classifier probability using Bayes' theorem again.

$$P_{BP}(c|E_{MAS}, E_F) = \frac{\left(\frac{P(c|E_{MAS})P(E_{MAS})}{\pi(c)}\right)\left(\frac{P(c|E_F)P(E_F)}{\pi(c)}\right)\pi(c)}{P(E_{MAS}, E_F)} \tag{4.7}$$

The known terms are substituted in.

$$P_{BP}(c|E_{MAS}, E_F) = \frac{\left(\frac{P_{MAS}(c|v,I)P(E_{MAS})}{\pi(c)}\right)\left(\frac{P_F(c|\boldsymbol{f})P(E_F)}{\pi(c)}\right)\pi(c)}{P(E_{MAS}, E_F)} \tag{4.8}$$

The class-independent terms $P(E_{MAS})$, $P(E_F)$ and $P(E_{MAS}, E_F)$ may be combined into a constant term. The final result is proportional to the product of the registration and the forest probabilities, divided by the class prior.

$$P_{BP}(c|E_{MAS}, E_F) = \left(\frac{P(E_{MAS})P(E_F)}{P(E_{MAS}, E_F)}\right)\frac{P_{MAS}(c|v,I)P_F(c|\boldsymbol{f})}{\pi(c)} \tag{4.9}$$

$$P_{BP}(c|E_{MAS}, E_F) \propto \frac{P_{MAS}(c|v,I)P_F(c|\boldsymbol{f})}{\pi(c)} \tag{4.10}$$

The constant of proportionality is removed by normalisation.

$$P_{BP}(c|E_{MAS}, E_F) = \frac{\frac{P_{MAS}(c|v,I)P_F(c|\boldsymbol{f})}{\pi(c)}}{\sum_{i=1}^{M}\frac{P_{MAS}(c_i|v,I)P_F(c_i|\boldsymbol{f})}{\pi(c_i)}} \tag{4.11}$$

4.3.5 Calibration of probabilities

We endeavour to correct for systematic inaccuracy in the probabilities, using the concept of calibration as described by Dawid [142].

The empirically measured probabilities, as observed in the training data, are quantised into B equally spaced bins plus a zero-probability bin. We deliberately choose $B = 14$ to equal the number of atlases. Fewer bins would not fully model the discrete set of atlas probabilities, and when calibrating the atlas result — which we do for comparison (see Tables 4.2 and 4.3) — this led to quantisation artefacts. We continue to use the same B for calibrating all methods since for the forest and hybrid approaches, experimentation with a greater number of bins gave no significant benefit.

The calibrated probability $P_{calibrated}(c|P(c|E_{MAS}, E_F)) \in b_i)$ for class c, given that the probability $P(c|E_{MAS}, E_F)$ falls within the range of bin b_i, $i = 1 \ldots B$, is as follows.

$$P_{calibrated}(c|P(c|E_{MAS}, E_F) \in b_i) = \frac{f(P(c|E_{MAS}, E_F) \in b_i | y = c)}{f(P(c|E_{MAS}, E_F) \in b_i)} \tag{4.12}$$

where $f(P(c|E_{MAS}, E_F) \in b_i)$ denotes the frequency of training voxels which have a predicted probability for c within the range of b_i. $f(P(c|E_{MAS}, E_F) \in b_i | y = c)$ denotes the frequency of voxels which have have a predicted probability for c within the range of b_i *and* have a ground truth label y equal to c. In an ideal classifier the predicted probability would already be equal to the calibrated probability.

Note that when performing calibration, we do a simple lookup of the calibrated value for the relevant bin. We do not interpolate between values, although interpolated line plots are shown in the calibration plots of Figure 4.5 for ease of interpretation.

4.3.6 Extraction of segmentation labels

The segmentation label C^* for a voxel v is the class with maximum probability in the calibrated class probability distribution for v.

$$C^* = \arg\max_c \; P_{calibrated}(c|E_{MAS}, E_F) \tag{4.13}$$

4.4 Evaluation

4.4.1 Evaluation measures

For the purpose of evaluation, we report three different metrics: mean Dice score, weighted mean Dice score, and error rate.

The **mean Dice score** [147, 148] is a common measure of segmentation accuracy (alongside the Jaccard index), which we compute according to the procedure of the MICCAI Grand Challenge against which we benchmark results.

- The mean is computed over 134 of the 207 ground truth classes, excluding regions whose labels do not appear in all datasets (as well as non-anatomical classes).
- Results are further broken down into cortical and non-cortical structures.
- The background class is not included.

We also compute the ***weighted* mean Dice score**, in exactly the same way as the mean Dice score, but weighting each class by its volume. We do this in order to show errors which may be masked by the DICE score since the mean Dice score weights classes equally regardless of volume (therefore magnifying error rates for small classes and suppressing those for larger classes).

Finally, we compute the **error rate** over the whole volume, *including* the background and excluded classes.

Mean Dice score

The Dice score for each class c is the ratio between the *intersection volume* and the *mean volume* of the ground truth segmentation Seg_{GT_c} and the automated segmentation Seg_{Auto_c}. The mean Dice score over all M classes is as follows.

$$\text{Mean DSC} = \frac{1}{M}\sum_{c=1}^{M}\left(\frac{2\,|Seg_{GT\_c}| \bigcap |Seg_{Auto\_c}|}{|Seg_{GT\_c}| + |Seg_{Auto\_c}|}\right) \tag{4.14}$$

A Dice score of one is achieved when the segmentations are identical. A Dice score of zero indicates that there is no overlap.

Weighted mean Dice score

We also report the *weighted* mean Dice score, where each class c is weighted by its volume V_c. This metric gives more importance to large classes.

$$\text{Weighted Mean DSC} = \frac{1}{\sum_{c=1}^{M} V_c} \sum_{c=1}^{M} V_c \left(\frac{2\,|Seg_{GT\_c}| \bigcap |Seg_{Auto\_c}|}{|Seg_{GT\_c}| + |Seg_{Auto\_c}|} \right) \tag{4.15}$$

Error rate

This is the percentage of voxels which are incorrectly classified. If y is the correct class label for a voxel v,

$$\text{Error Rate} = 100\% \times \frac{1}{V} \sum_{v=1}^{V} C_v^* \neq y_v \tag{4.16}$$

4.4.2 Results

4.4.2.1 Quantitative results

Mean Dice scores are shown in Table 4.2. Dice scores are shown for the winner of the MICCAI Grand Challenge 2012 (Wang *et al.* [149]), a group from the University of Pennsylvania who used a combination of FLIRT [150, 151, 152] for affine registration and ANTS-Syn [6] for non-rigid registration, followed by label fusion by image similarity based local weighted voting. Results are also shown for the algorithm of Zikic *et al.* [21] mentioned previously, the *atlas forest*. See Table 4.3 for results of the weighted mean Dice score and error rate.

It is evident that the decision forest alone is a much poorer classifier than the multi-atlas registration algorithm. However, we seek two classifiers which yield uncorrelated information, so individual performance is less interesting. In fact, the rankings of the decision forest classifiers are reversed when they are combined with registration (with the caveat that A and B are swapped if the error rate is used as the metric).

Classifier A performs best when the Bayesian product is used. Classifiers B and C achieve best results when combined by averaging. Since the quantitative improvements in results are small, we also performed significance tests regarding

Method	Mean Dice Score		
	Overall	**Cortical**	**Non-Cortical**
Wang *et al.*	0.7654	0.7388	0.8377
Zikic *et al.*	0.7366	0.7104	0.8081
MAS	0.730 *(0.727)*	0.707 *(0.703)*	0.793 *(0.791)*
Classifier A			
Forest	0.389 *(0.028)*	0.321 *(0.000)*	0.574 *(0.104)*
Average	0.734 *(0.666)*	0.711 *(0.633)*	0.797 *(0.754)*
Bayesian Product	0.743 *(0.719)*	0.719 *(0.698)*	0.807 *(0.777)*
Classifier B			
Forest	0.595 *(0.190)*	0.568 *(0.120)*	0.669 *(0.381)*
Average	0.739 *(0.715)*	0.716 *(0.692)*	0.803 *(0.778)*
Bayesian Product	0.733 *(0.700)*	0.708 *(0.675)*	0.799 *(0.769)*
Classifier C			
Forest	0.615 *(0.162)*	0.583 *(0.074)*	0.700 *(0.401)*
Average	0.732 *(0.714)*	0.708 *(0.690)*	0.797 *(0.779)*
Bayesian Product	0.716 *(0.664)*	0.690 *(0.636)*	0.786 *(0.739)*

Table 4.2: Mean Dice score results for different brain segmentation classifiers. Figures are computed over the 20 test datasets, and are further broken down into the mean Dice score over cortical and non-cortical structures. The numbers in brackets indicate the raw (uncalibrated) scores. The best combination method for each classifier is highlighted in grey.

the number of datasets that improved in the hybrid classifiers compared to the MAS method (see Table 4.4). From this table, it can be seen that for classifiers A and B there was a consistent improvement across the board. The probability for classifier C is lower, but would still be significant at the level $p = 0.01$.

4.4.2.2 Qualitative results

The qualitative improvement made by the combined classifiers over and above that of registration alone can be appreciated in the images in Figure 4.3. Further intuition can be gained by viewing Figure 4.4 which shows the the decision forest classifier segmentations for the same slice as in Figure 4.3. The features for classifiers A and B are all *relative* features which measure intensities relative to the

Method	Weighted Mean Dice Score			Error Rate (%)
	Overall	Cortical	Non-Cortical	
MAS	0.829 *(0.828)*	0.741 *(0.737)*	0.910 *(0.911)*	5.850 *(5.928)*
Classifier A				
Forest	0.473 *(0.187)*	0.753 *(0.000)*	0.924 *(0.361)*	18.829 *(25.257)*
Average	0.833 *(0.795)*	0.746 *(0.671)*	0.912 *(0.909)*	5.743 *(7.835)*
Bayesian Product	0.842 *(0.828)*	0.753 *(0.738)*	0.924 *(0.912)*	5.560 *(6.475)*
Classifier B				
Forest	0.720 *(0.434)*	0.597 *(0.133)*	0.833 *(0.712)*	10.098 *(18.489)*
Average	0.840 *(0.826)*	0.751 *(0.727)*	0.922 *(0.917)*	5.424 *(6.077)*
Bayesian Product	0.839 *(0.817)*	0.746 *(0.716)*	0.925 *(0.911)*	5.748 *(7.409)*
Classifier C				
Forest	0.727 *(0.407)*	0.603 *(0.082)*	0.842 *(0.708)*	9.546 *(18.518)*
Average	0.831 *(0.820)*	0.742 *(0.722)*	0.914 *(0.910)*	5.799 *(6.270)*
Bayesian Product	0.822 *(0.762)*	0.729 *(0.676)*	0.908 *(0.842)*	6.631 *(9.676)*

Table 4.3: Weighted mean Dice score and the error rate results for different brain segmentation classifiers. Figures are computed over the 20 test datasets. The weighted mean Dice score figures are further broken down into the weighted mean Dice score over cortical and non-cortical structures. The numbers in brackets indicate the raw (uncalibrated) scores. The best combination method for each classifier is highlighted in grey.

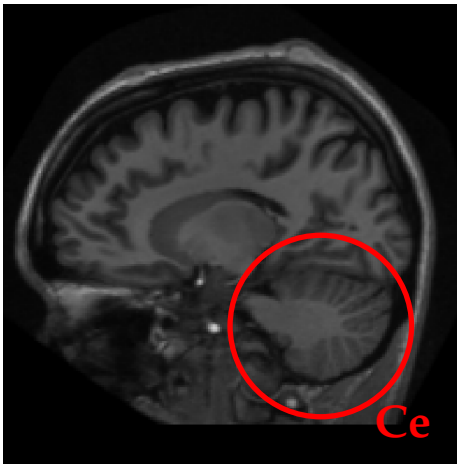
voxel of interest. Hence these classifiers pick out the fine detail of the boundaries between tissues. Since classifier A uses very local features (displacement $\leq$ 15mm), the segmentation is especially fragmented. However the locality of the classifier also means that it has the least correlation with the spatially-driven registration algorithm and so it might be expected that the Bayesian product would be a suitable method of combination, with its assumption of classifier independence. Classifiers B and C use significantly larger feature displacements (displacement $\leq$ 50mm) and so they are more spatially aware, hence averaging results with the registration algorithm works well.

Method	Mean Dice	No. Improved Datasets	p
Classifier A (Bayesian)	0.743	20/20	$9.54x10^{-7}$
Classifier B (Average)	0.739	20/20	$9.54x10^{-7}$
Classifier C (Average)	0.732	16/20*	0.0059

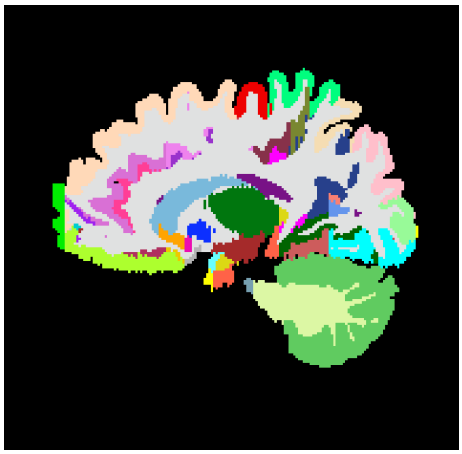
Table 4.4: Significance test results for the hybrid brain segmentation classifier compared to multi-atlas segmentation alone, based on how many datasets improved. The binomial probability p of this number of datasets (or more) improving by chance is given, using the binomial probability distribution (n = 20, k = number of improved datasets, p = 0.5). *Two datasets stayed the same, and these are considered to have not improved.

4.4.2.3 Visualisation of calibration plots

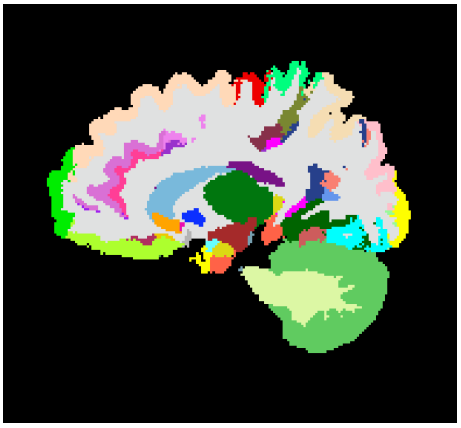
Calibration of probabilities clearly contributes significantly to the success of this method. Calibration is effectively a simple machine learning step i.e. *given a predicted class probability, predict the true class probability*. We show some contrasting example calibration plots in Figure 4.5.



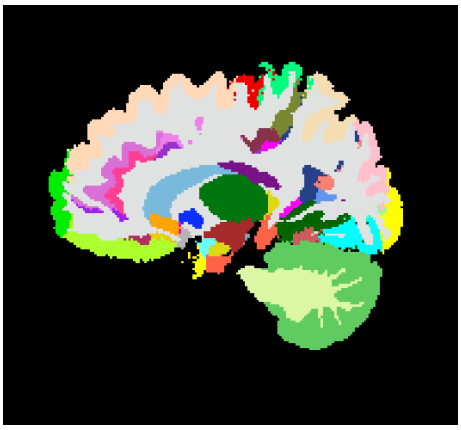
(a) Sagittal Slice



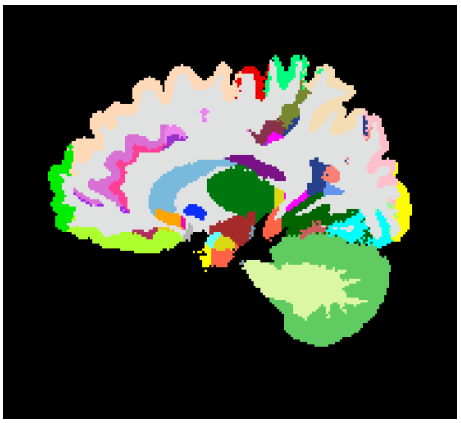
(b) Ground Truth



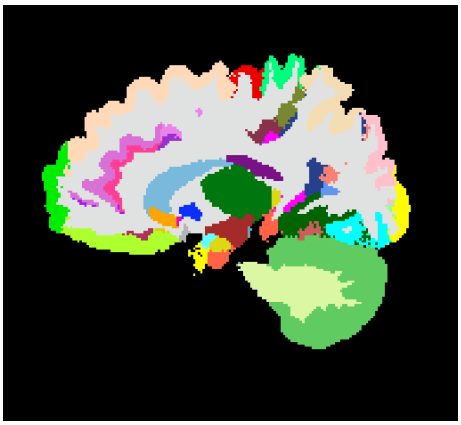
(c) Multi-Atlas Registration



(d) Classifier A, Bayesian Product

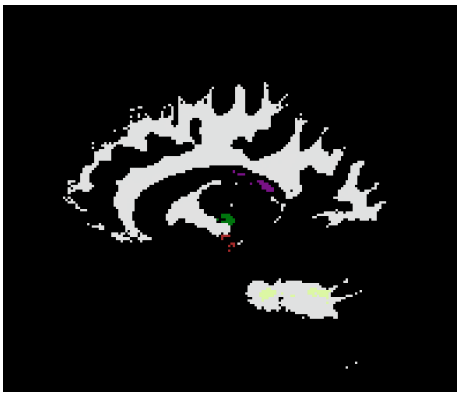


(e) Classifier B, Arithmetic Mean

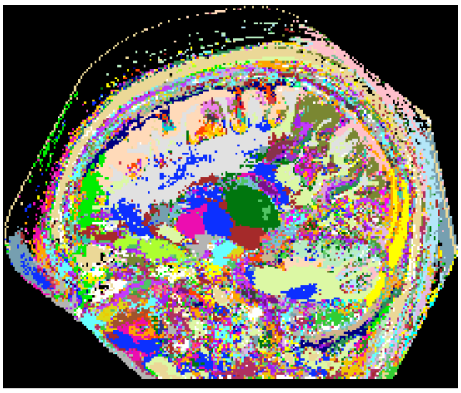


(f) Classifier C, Arithmetic Mean

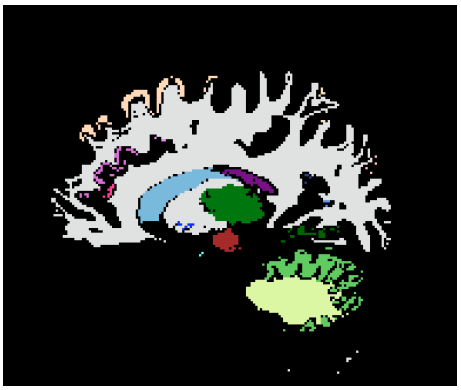
Figure 4.3: Segmentation label results for a slice in an example dataset. The best combination method is shown for each classifier. The difference is particularly apparent in the white matter of the cerebellum (Ce), which is the pale green inside the darker green of the cerebellum grey matter cortex. [TMVS Dataset ID: 9954]



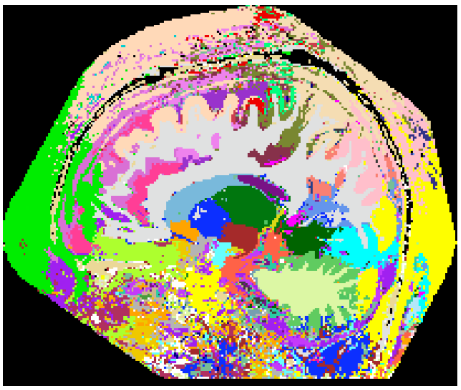
(a) Classifier A, $\arg\max_c P_F(c|\boldsymbol{f})$



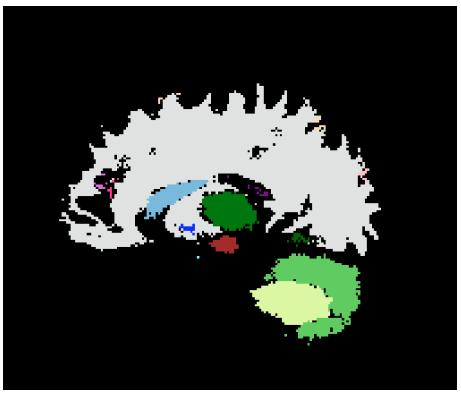
(b) Classifier A, $\arg\max_c \frac{P_F(c|\boldsymbol{f})}{\pi(c)}$



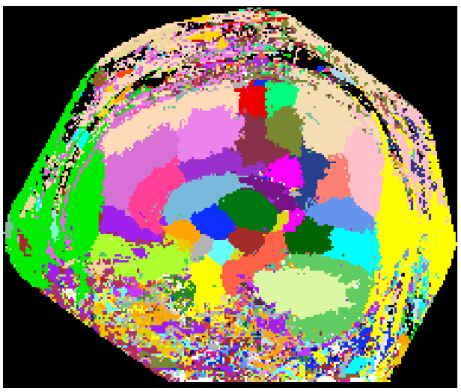
(c) Classifier B, $\arg\max_c P_F(c|\boldsymbol{f})$



(d) Classifier B, $\arg\max_c \frac{P_F(c|\boldsymbol{f})}{\pi(c)}$

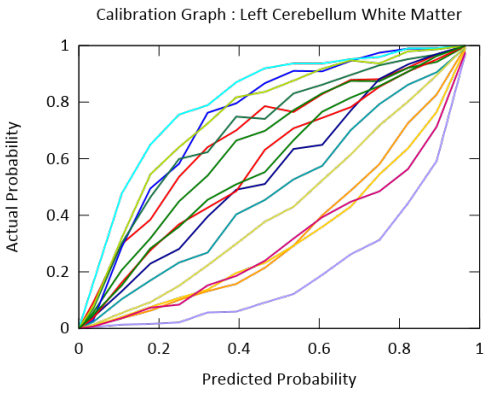


(e) Classifier C, $\arg\max_c P_F(c|\boldsymbol{f})$

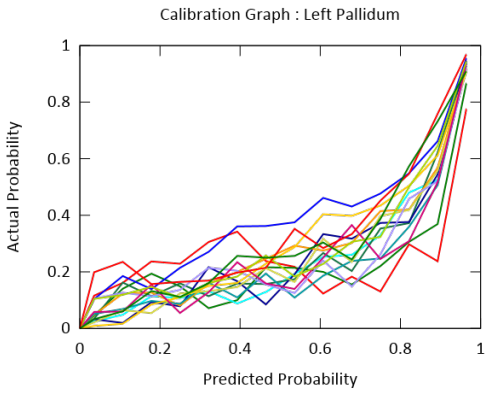


(f) Classifier C, $\arg\max_c \frac{P_F(c|\boldsymbol{f})}{\pi(c)}$

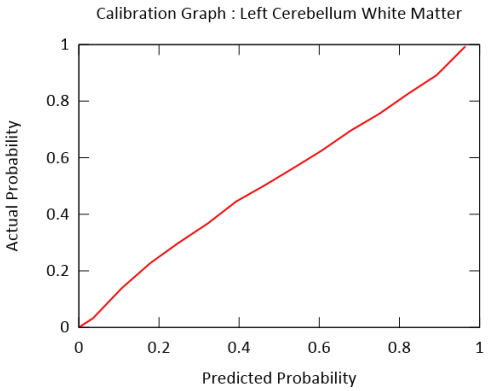
Figure 4.4: Forest-only segmentation results for the sagittal slice of Figure 4.3, illustrating the difference between the three alternative random forest classifiers. Left: Segmentation labels given by the forest posterior probabilities. Right: Segmentation labels given by the forest posterior probabilities *divided* by the class prior probabilities (effectively uniform prior probabilities). The right-hand images have been included to give insight into the probability distributions of small regions which have low prior probabilities. [TMVS Dataset ID: 9954]



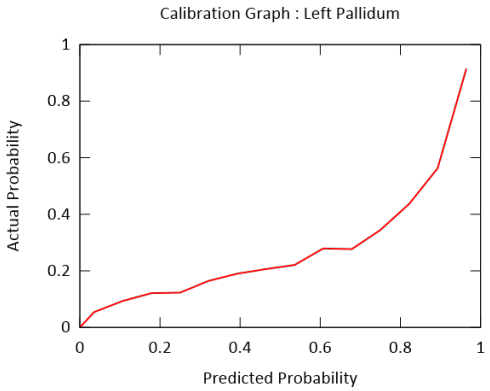
(a) Left Cerebellum White Matter



(b) Left Pallidum



(c) Left Cerebellum White Matter



(d) Left Pallidum

Figure 4.5: Calibration plots for two example brain structures, using the calibrated Bayesian product result of classifier A. The x-axes show the predicted probability according to the classifier, and the vertical axes show the empirically measured true probability. In a well calibrated classifier, these should be equal (i.e. a perfect diagonal $x = y$). Figures a) and b) show the datasets individually (each coloured line represents a different dataset). Figures c) and d) show the aggregated figures across all datasets: these numbers are the ones used for the purpose of performing calibration. We chose to use empirical results rather than fitting a function. Consequently, a few classes have bumpy curves such as that exhibited in d).

Classifier	Training Time	Detection Time	Selective Detection Time
A	2 hours	6 mins	2 mins
B	5 hours	13 mins	4 mins
C	13 hours	40 mins	8 mins

Table 4.5: Run times for training and detection for three alternative brain segmentation random forest classifiers. The 'Selective Detection Time' refers to the detection time when voxels are omitted which have been classified by the registration algorithm as background with a probability of 1.0. Note that the registration step adds a further 5 minutes to the run time.

4.4.3 Run times: A trick and a trade-off

4.4.3.1 Trick for reducing the detection time

Classifier A has the fastest run time, both for training and for detection, since fewer features are used (F_T = 200) and these features are effectively simple voxel intensity look-ups. Classifiers B and C use a greater number of features (F_T = 1000) and, using the integral volume representation, each feature value requires 8 voxel look-ups.

There is a further trick for speeding up the detection time by selectively running detection over a subset of voxels in the volume. We notice that for the Bayesian product result, any class for which either P_{MAS} or P_F is zero must also have a Bayesian product probability $P_B P$ of zero (note we do not calibrate voxels with a probability of zero). Therefore, after running registration, we detect all voxels for which background is given a probability of 1.0, and we can automatically classify these voxels as background. This gives a significant time saving, since background accounts for approximately 70% of each dataset (the majority of which are correctly classified as certain by the registration algorithm). An experiment on the training data showed that only 0.3% of foreground voxels were classified as certain background by the registration algorithm, so we propose that it is reasonable to use this trick for the averaging combination method also.

Run times are summarised in Table 4.5. The training time is a one-off cost that does not need to be repeated for each novel dataset.

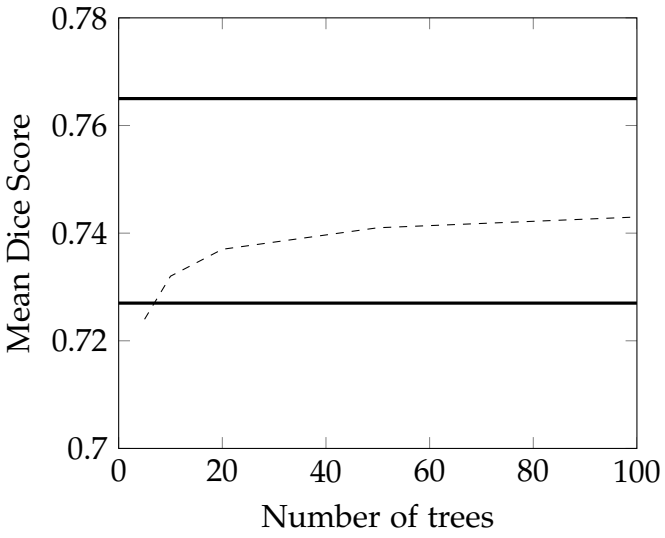


Figure 4.6: Graph showing the relationship between the number of trees in the forest classifier, and the mean Dice score. Figures are computed for the calibrated Bayesian product result of classifier A. The solid lines indicate (upper) the result of Wang *et al.* and (lower) the result of multi-atlas registration. The classification time for the forest scales linearly with run time.

4.4.3.2 Trade-off between accuracy and run time

We also investigated the effect of reducing the size of the forest. A graph showing the trade-off between forest size and accuracy can be seen in Figure 4.6. We compare the Bayesian product obtained with classifier A to the accuracy of registration alone, and also that of Wang *et al.* . The forest size could be halved from 100 to 50 tress with little loss in accuracy, thus halving the run times (since there is a linear relationship between number of trees and run time, at least in our serial implementation).

The test computer has a dual-processor, 24-core Intel Xeon CPU with clock speeds of 2.40GHx and 2.39GHz, and 32GB RAM. On similar hardware, for the algorithm of Wang *et al.* [149], the main computation time was for image registration which takes approximately 20 hours for each pair of images on a 2G HZ CPU (i.e. 15 datasets $\times$ 20 hours in total, per novel dataset). Hence the whole process of registration is much slower. For comparison, the algorithm of Zikic *et al.* takes approximately 4 minutes in run time which is a little faster. This is because we do $A = 15$ registrations whereas Zikic do a single registration of the novel to the reference atlas, also using an efficient registration algorithm [153] (this step takes 30 seconds, so is slightly slower than our single registration time of 20 seconds).

4.4.4 A note on Bayes and the use of class priors

The results presented for the Bayesian product in this chapter were performed by working with the posterior probabilities of each algorithm.

During experimentation, slightly better results were achieved by using the re-weighting technique as was used in anatomical landmark detection, see equation 2.13, and by Zikic [21] *in advance of* tree training. This technique imposes uniform class prior probabilities at the tree-level of the forest step, and hence allows simple multiplication of the forest and registration results (without division by the class priors). To do this, we redefine equation 4.2 as follows, where $f_R(c)$ is the frequency of training samples from class c at the root node.

$$P_T(c|v) = \frac{\frac{f_{Lv}(c)}{f_R(c)}}{\sum_{i=0}^{M} \frac{f_{Lv}(c_i)}{f_R(c_i)}} \tag{4.17}$$

Later, when substituting in the known terms (equation 4.8), we have the registration algorithm which uses the "true" priors and the forest algorithm which uses the constant (uniform) prior, $\frac{1}{M+1}$.

$$P_{BP}(c|E_{MAS}, E_F) = \frac{\left(\frac{P_{MAS}(c|v)P(E_{MAS})}{\pi(c)}\right)\left((M+1) \times P_F(c|\boldsymbol{f})P(E_F)\right)\pi(c)}{P(E_{MAS}, E_F)} \tag{4.18}$$

Then, if all constant terms are gathered together, the final result is proportional to the product of the registration and the forest probabilities.

$$P_{BP}(c|E_{MAS}, E_F) = \left(\frac{(M+1) \times P(E_{MAS})P(E_F)}{P(E_{MAS}, E_F)}\right) P_{MAS}(c|v, I)P_F(c|\boldsymbol{f}) \tag{4.19}$$

$$P_{BP}(c|E_{MAS}, E_F) \propto P_{MAS}(c|v, I)P_F(c|\boldsymbol{f}) \tag{4.20}$$

This method gives results as shown in Tables 4.6 and 4.7.

We make the observation that the re-weighting method only partially imposes the desired priors (in this case uniform priors). Take the extreme case, where

Method	Mean Dice Score		
	Overall	Cortical	Non-Cortical
Classifier A	0.744 *(0.734)*	0.720 *(0.712)*	0.811 *(0.793)*
Classifier B	0.741 *(0.729)*	0.717 *(0.706)*	0.806 *(0.791)*
Classifier C	0.733 *(0.710)*	0.707 *(0.685)*	0.801 *(0.779)*

Table 4.6: Mean Dice score results for hybrid brain segmentation classifiers using *tree*-level Bayesian priors (uniform prior re-weighting).

Method	Weighted Mean Dice Score			Error Rate (%)
	Overall	Cortical	Non-Cortical	
Classifier A	0.842 *(0.838)*	0.753 *(0.749)*	0.923 *(0.920)*	5.487 *(5.809)*
Classifier B	0.841 *(0.838)*	0.751 *(0.745)*	0.923 *(0.924)*	5.445 *(5.890)*
Classifier C	0.831 *(0.814)*	0.739 *(0.722)*	0.916 *(0.899)*	6.009 *(7.091)*

Table 4.7: Weighted mean Dice score and error rate results for hybrid brain segmentation classifiers using *tree-level* Bayesian priors (uniform prior re-weighting).

every end leaf is a pure one containing samples from a single class. Re-weighting has no effect since the probabilities in a leaf are subsequently normalised to sum to one. However, the forest may yield an ensemble of pure leaves from different classes, which gives a non-pure forest posterior probability distribution. Re-weighting at the level of the forest then has an effect. The proper method of multiplying the posterior probabilities and dividing by the class prior is equivalent to re-weighting at the level of the forest.

The fact that this method gives better performance, particularly in terms of error rate, likely stems from the weak dependency between the registration and forest classifiers. The biggest improvement in mean Dice score is seen for classifier B, which fits in with the fact that this classifier has both spatial and local intensity awareness, and hence neither the mean nor the true Bayesian product are a good fit.

4.4.5 Are the improvements in results meaningful?

The improvements that we have been demonstrating are qualitatively visible, particularly at the boundary between grey and white matter regions, but they

are very small in numerical terms. This small spread in numerical results was also a feature of the MICCAI Grand Challenge, and of papers published since by Zikic *et al.* [21], Cardoso *et al.* [154] and Ledig *et al.* [122] who achieved mean Dice scores of 0.737, 0.755 and 0.772 respectively (Ledig *et al.* surpassed the challenge winner by a narrow margin of 0.007). Figure 4.7 shows the table of results against which we are comparing. It can be seen that all 25 entrants placed in a range of 0.06, of which 20 entrants were within 0.03.

It may be that we are nearing the limits of achievable accuracy, despite the low Dice scores. No figures are available for inter-observer error, but it can be seen visually that the boundaries are quite indistinct and so would be difficult to consistently pick out and follow. Figure 4.8 illustrates horizontal striping artefacts in the ground truth. This is evidence that the ground truth collection is not consistent from slice to coronal slice. We note further, that for small or narrow structures, overlap errors which are small in absolute terms will be heavily penalised.

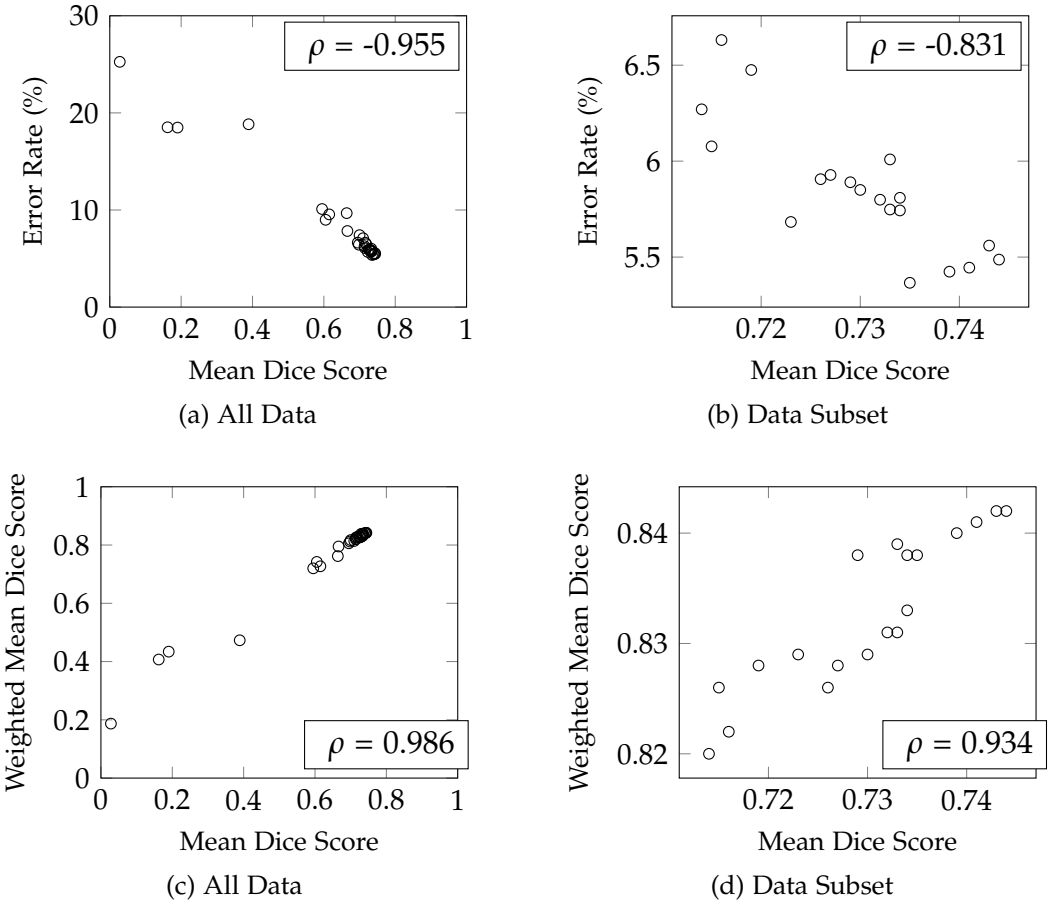
In order to explore the significance of the small differences in performance, our analysis is extended to an examination of the correlation between different error metrics. If all permutations of the algorithm are considered, including both calibrated and uncalibrated versions, we now have performance metrics for 32 algorithms. In Figure 4.9, the mean Dice scores of the algorithms are plotted against each of the other metrics. Correlation coefficients are displayed on the graphs. Spearman's ranking correlation coefficient is computed, to see how well the metrics agree on the relative performance of the many algorithms that we have reported.

In the graphs in Figure 4.9, it can be seen that the metrics are generally strongly correlated, and this lends some confidence that in our work we have been demonstrating real improvements quantitatively as well as qualitatively, irrespective of the measure used. There is slightly weaker correlation at the top end of the scale, particularly when comparing Dice score with error rate. Differences are due to the former rating each *voxel* in the volume equally, and the latter rating each *region* equally (regardless of relative size).

Overall Rank †	Team Name	Mean DSC Overall	Mean DSC Cortical	Mean DSC Non-Cortical
1	PICSL_BC	0.7654	0.7388	0.8377
2	NonLocalSTAPLE	0.7581	0.7318	0.8296
3	MALP_EM	0.7576	0.7328	0.8252
4	PICSL_Joint	0.7499	0.7216	0.8271
	Classifier A, Bayesian	**0.743**	**0.719**	**0.807**
5	MAPER	0.7413	0.7144	0.8144
6	STEPS	0.7372	0.7107	0.8095
7	SpatialSTAPLE	0.7372	0.7093	0.8130
8	CIS_JHU	0.7357	0.7131	0.7971
	Murphy MAS	**0.7346**	**0.7183**	**0.7807**
9	CRL_Weighted_STAPLE_ANTS+Baloo	0.7344	0.7122	0.7950
10	CRL_Weighted_STAPLE_ANTS	0.7308	0.7066	0.7966
11	CRL_STAPLE_ANTS+Baloo	0.7290	0.7064	0.7919
12	CRL_STAPLE_ANTS	0.7280	0.7033	0.7951
13	CRL_Probabilistic_STAPLE_ANTS+Baloo	0.7251	0.7009	0.7911
14	CRL_MV_ANTS+Baloo	0.7247	0.6966	0.8012
15	CRL_MV_ANTS	0.7243	0.6951	0.8035
16	DISPATCH	0.7243	0.6965	0.8000
17	CRL_Probabilistic_STAPLE_ANTS	0.7223	0.6972	0.7907
18	SBIA_SimRank+NormMS+WtROI	0.7212	0.6940	0.7953
19	SBIA_BrainROIMaps_MV_IntCorr	0.7193	0.6933	0.7904
20	SBIA_BrainROIMaps_JaccDet_IntCorr	0.7186	0.6913	0.7927
21	BIC-IPL-HR	0.7173	0.6888	0.7948
22	SBIA_SimMSVoting	0.7172	0.6898	0.7918
23	UNC-NIRAL	0.7171	0.6869	0.7992
24	SBIA_SimRank+NormMS	0.7162	0.6884	0.7919
25	BIC-IPL	0.7107	0.6829	0.7864

Figure 4.7: Table of results from the 2012 MICCAI Grand Challenge on multi-atlas labelling [1]. Results are inserted for Murphy's original MAS algorithm [26] (using expectation maximisation rather than calibration as a post-processing step) and for the best performing hybrid classifier A (Bayesian Product).

Figure 4.8: An axial slice from the dataset used in Figures 4.3 and 4.4, showing horizontal striping artefacts resulting from manual collection of ground truth in the coronal plane. Note that there are some legitimate horizontal boundaries between structures due to the fact that some boundaries were defined by a coronal plane aligned with a particular defined landmark. [TMVS Dataset ID: 9954]



(a) All Data (b) Data Subset

(c) All Data (d) Data Subset

Figure 4.9: Scatter graphs showing mean Dice score plotted against error rate and against weighted mean Dice score. Left: All algorithms in this chapter. Right: Subset of algorithms with scores between 0.714 and 0.744. Spearman's rank correlation coefficient, ρ, is shown for each graph.

4.5 Discussion

4.5.1 A summary of our contribution

We have demonstrated that there is value in combining classifiers with complementary strengths to give a superior combined classifier. By taking two efficient algorithms (multi-atlas registration and random classification forest) and combining them using a simple operation, a method of automatic brain segmentation is produced which is fast relative to the state of the art, with only modest shortfall in accuracy.

We tried both intensity and gradient orientation features for the random forest classifier. The gradient orientation features performed best, turning out to be particularly good at spatially locating a voxel. These features learn aggregated information about the local neighbourhood but are blind to the particular intensity of the voxel of interest, making these features good for approximate location of a voxel but poor for discrimination between tissues. However, this random forest classifier was slow to train and detect with, since gradient orientation features are computationally expensive, and did not perform better than the MAS classifier.

The short-range relative intensity features gave the best *combined* classifier performance. The preferred method of combination depended on the degree of dependency between the classifiers. We discussed the fact that application of the Bayesian product priors at the level of the tree rather than at the level of the forest (thus only weakly imposing the priors) gave slightly better results for all three classifiers. This probably indicates that neither averaging nor multiplying the classifier results is ideal. We discuss the conditional independence assumption further in section 4.5.2 and potential alternative combination methods in section 4.6.2.

4.5.2 The importance of independence when combining evidence

In the previous chapters 2 and 3 on anatomical landmark detection, we stressed the importance of *randomness* in enabling an ensemble classifier such as a decision forest to have good generalisation properties to unseen data. By introducing diversity amongst the cohort of classifiers (through the use of bagging [49] and random feature subspaces [50]), we aimed to train a classifier which achieves a good bias-variance trade-off, as discussed in section 3.9.1. The role of diversity

in random forests is not well understood [155, 139] except as being an essential component. However, Gashler *et al.* [156] found that heterogeneity amongst the tree algorithms, by mixing two types of trees trained using different nodal split selection criteria (entropy and *mean margins*), conferred robustness to irrelevant attributes, and in particular fewer trees were required to achieve the same level of performance.

In this chapter on brain segmentation, we have gone further and mixed two highly heterogeneous classifiers which have very different domain models, although both yield probabilistic outputs. On this basis, we followed the lead of Kittler [138] in assuming conditional independence between the classifiers. The resulting Bayesian product combination has indeed outperformed simple averaging for the best performing hybrid classifier.

Nonetheless, we have produced a comprehensive set of results showing that it is important to choose features for the decision forest classifier which maximise the difference between the forest and the MAS algorithm. Only in the circumstances where these two are sufficiently different is the conditional independence assumption sufficiently justified, or indeed is there benefit in combining classifiers at all.

4.5.3 Why does calibration of probabilities make such a significant difference?

The simple empirical calibration step makes a great improvement to the results. In particular, for the forest classifier, we would expect the output probability distributions to be close to the empirically observed probability distributions, since the leaf class distributions were generated directly from the training data frequencies.

Upon examination of the forest calibration plots (not included in this thesis), it was seen that the class probabilities were consistently underestimated compared to the true values — except for where the true probability is zero. Inspection of the leaf class probability distributions revealed that each leaf contains in the order of tens of classes, each with a small probability, showing that classes are being only weakly separated within each decision tree.

We tentatively suggest that the forest ensemble is actually modelling a more complex hypothesis space than that of which a single tree is capable. This would tie in with the theory of Monteith *et al.* [157] who proposed that the power of ensembles comes, at least in part, from the enriched hypothesis space and more

general bias that can be provided by a combination of models. In this case, the *combination* of T leaves into which a voxel sample falls may be significant — and whilst the nominal averaged probability of a class across a number of leaves may be low, in fact the probability of the class *given* that it has reached all of these T leaves may be high. Further analysis is required to ascertain the truth of the matter.

4.6 Future work

4.6.1 Validation on more data: Quantity, diversity and clinical relevance

The number of training datasets (15) was very small for a machine learning method, and we anticipate that a greater number of training datasets would make a significant improvement to performance.

Regarding real world application, it would be helpful to perform validation on more diverse datasets. The datasets used in this chapter have relatively homogeneous intensity distributions, since they are generated with the same scan protocol and have had the same extensive pre-processing steps applied (see section 4.3.1), in order to produce high quality research data. The next step is to ensure that our algorithm is robust to different protocols and data quality. Perhaps some explicit modeling of tissue intensities (e.g. as a mixture of Gaussians) might be used to devise a more robust intensity normalisation scheme.

Finally, it is important to point out that brain segmentation is most useful on pathological cases, where the brain can be quite distorted due to processes such as tissue atrophy or plaque formation. Many cases in this study were of young, healthy individuals, and we did not have access to clinical diagnoses. It would be useful to perform explicit, quantitative validation on clinical cases.

4.6.2 Moving to more complex combination methods

Empirical calibration of the probabilities is effectively a simple machine learning step. In view of the significant improvement that calibration made, the logical progression would be to replace both calibration and combination steps with an explicit machine learning method. This would remove the need to make assumptions about the independence of classifiers and would allow more complex,

data-driven relationships to be modelled.

Stacked generalisation [158] refers to the stacking of a machine learning method on top of the base classifiers, by taking the base classifier outputs as inputs and attempting to learn the correct outputs. The simplest approach would be softmax regression, alternatively known as multinomial logistic regression, which is the generalisation of logistic regression to multiclass problems.

Chapter 5

Arterial tree tracking from anatomical landmarks

Abstract

In this chapter, we prototype an application of anatomical landmarks: centre line tracking of the major arteries in MRA scans. We first optimise the choice of intensity features for detection of vascular landmarks in angiography scans. For the vessel tracking, we define a graph representing the standard arterial system, with nodes denoting landmarks at vascular branch points and termini and edges denoting the connecting vessels. Given a novel scan, we identify which vessels correspond to the positively detected landmarks. The vessels are then tracked one by one using a pre-existing shortest path tracking algorithm, which we augment with contextual information, in the form of vessel enhancement and a vascular atlas. Vessel enhancement is achieved by morphological filtering of the scan at a scale corresponding to the expected vessel diameter, and tracking on this filtered volume. A vascular atlas is created by registering the ground truth centre lines together, and then mapping this to the novel volume, to make a prediction about the expected path of the volume; the atlas comprises part of the cost function when tracking. Contextual information is shown to improve results, even when tracking from manually placed landmarks. Tracking from detected landmarks is only a little worse than from manually placed landmarks, and demonstrates good potential for future clinical application.

5.1 Synopsis

In this chapter we

- (5.3.1) Investigate intensity feature parameter choices for landmarks in cardiac Computed Tomography Angiography (CTA) scans, showing that simply using *relative* rather than absolute intensity features in CT data gives a better failure mode for vascular landmarks, in the context of vessel tracking.
- (5.4) Define a set of vascular landmarks which describe the standard arterial tree for the whole body.
- (5.5.3 and 5.5.4) Describe how contextual information may be used to inform a vessel tracking algorithm, specifically through knowledge of the typical diameter of a vessel (for pre-processing with the appropriately sized vessel filter) and spatial path (for selecting an atlas of the relevant vessel to inform the tracking cost function).
- (5.5.3.3 and 5.7.2) Present a comparison of the morphological top-hat transform and the Frangi vesselness Hessian-based filter, for the purpose of vessel enhancement filtering. Explain why the former gives better practical results in the context of vessel tracking.
- (5.5.6) Using *manually* placed landmark points, demonstrate vessel tracking in MRA scans, showing that contextual information gives significant benefit.
- (5.6.2) Using *automatically* detected landmark points, demonstrate vessel tracking in the same MRA scans, showing that the fully automatic vessel tracking method gives reasonable results, which are only a little worse than the manually landmarked results.

5.2 Introduction

5.2.1 Problem description

This chapter examines the problem of automatically identifying and tracking the major arteries which are present in an MRA scan (of any acquisition region). Vessel centre line tracking has clinical applications in the visualisation and monitoring of vascular disease such as thrombosis and atherosclerosis, and in the planning and monitoring of surgical operations. Once the path of a vessel is known, it may be visualised in the curved plane of the path, enabling navigation along the vessel lumen to conveniently locate or examine pathology. The centre line may be used to initialise subsequent segmentation e.g. as in [159, 160], giving automatic measurement of the vessel diameter, which is a common clinical measurement of interest. Ultimately more advanced analysis may be spawned such as aneurysm analysis [161] or texture-based plaque classification [162].

Manual vessel tracking is time-consuming, and may be problematic in the presence of noise, low vessel contrast, other bright distracting structures in the neighbourhood, or variation which may be pathological or anatomical in nature. We aim to develop a fully automatic method which can identify and track the vessels which are present in a given MRA scan.

5.2.2 Prior art

There have been various reviews of vessel extraction techniques [163, 164, 165, 166, 167, 168, 169], most recently that of Lesage *et al.* [169]. The definition of vessel "extraction" may range from simply enhancing or segmenting vessel pixels, to identifying the vascular topology (usually assumed to be a tree structure), to labelling of the vessels with their anatomical names.

We do not attempt a complete review of the literature, but give an overview of the approaches available, with emphasis on the aspects in which we are interested. The review is divided into two parts, first considering methods of enhancing vascular structures and secondly considering approaches to extracting and labelling the vascular topology.

5.2.2.1 Vessel enhancement

Distinguishing vessels from other anatomical structures — we will hereafter refer to non-vessel regions of the medical scan as *background* — is important for all

approaches to vessel tracking and segmentation. Vessel enhancement filtering techniques focus on the small size (relative to other organs), tubular shape and brightness of vessels. Indeed, simple thresholding of the data may provide a reasonable separation of vessels from background, since vessels are usually one of the brighter objects in a medical image e.g. in contrast-enhanced, time-of-flight or phase contrast MR scans. This is not always the case. In modalities such as black-blood imaging the vessels are dark structures and the filter mathematics must be inverted.

Morphological filters [170] are made up of erosion, dilation, opening and closing operations according to shapes called *structuring elements* (SEs), yielding a non-linear filtering result. These operations originated for use with binary images, and adaptations for grey-scale data have subsequently evolved. One filter commonly used for enhancement of small bright objects is the top-hat transform which consists of the difference between an image and its opened version. Another example is the hit-or-miss transform (HMT). Introduced by Matheron and Serra [171, 172], the HMT uses the erosion operator and a pair of disjoint SEs A and B. The result is the set of positions, where A fits in the foreground and B wholly misses it. A number of extensions of the HMT to grey-level data have been proposed [173, 174, 175], including an extension tailored for vessel enhancement in MR angiography data by Naegel *et al.* [176, 177]. By contrast to these localised filters, Merveille *et al.* [178] suggested a morphological path opening operator. Path operators process lines of connected voxels, where connections are made according to some adjacency criteria. During path opening, paths with a length less than some L are rejected. In the implementation of Merveille *et al.* , path opening was applied in seven orientations according to adjacencies defined by conical SEs corresponding to the three vectors of the orthogonal basis x, y, z and the four principal diagonals. The responses were ranked, and it was found that tubular structures always have a response ranked no lower than third. The filter response was then given by the top-ranked response minus the fourth-ranked response.

Curvature-based filters use analysis of the second order derivatives of the local voxel intensities to elicit a tubularity or "lineness" response. The basic idea is that curvature will be low-to-zero in the direction of the vessel (depending on the gradient of the vessel contrast), with high curvature perpendicular to it. Du *et al.* [179] applied the second order differential operator in all 13 directions possible in a 3x3x3 voxel cube, and chose as the vessel direction that which had maximum difference compared to the eight directions perpendicular to

it. However, this filter does not distinguish lines from surfaces. Later authors [180, 181, 182, 183, 184, 185, 186, 187] analysed the eigenvalues of the Hessian matrix (i.e. matrix of second-order partial derivatives). Possibly the most well-known is the vesselness filter of Frangi *et al.* [182]. This filter uses scale-space theory [188, 189, 190] to achieve multi-scale vessel enhancement. Three different ratios of the three eigenvalues (identified by their magnitude, with the smallest eigenvalue corresponding to the eigenvector in the vessel direction) were used to differentiate blobs from lines from surfaces, and to suppress noise in the image. Chapman [191] compared the filters of Du *et al.* and Frangi *et al.* and found performance to be similar.

Some authors have used the idea of gradient vectors or the diffusion thereof. Bauer and Bischof [192] suggested a modified vesselness filter, where the isotropic Gaussian smoothing step (to achieve linear scale-space) is replaced by an anisotropic diffusion process [193]. Regions with small vectors are smoothed much more than points with large vectors to give a smooth (slowly varying) result where initial vector magnitudes are small, while preserving vectors with high magnitude. At the centres of tubular objects, the resultant gradient vector field (GVF) has the same properties as if gradients had been smoothed at the appropriate (multiple) scales, thus Hessian analysis can be performed directly on the GVF. Moving on to detection of the vessel boundary, Vasilevskiy and Siddiqi [159] developed an active contour model based on the idea that vessels are points of high inward gradient flux and that vessel contours sit at the position of highest rate of change of flux. In *optimally oriented flux*, Law and Chung [194] did likewise, but coupled this with computation of the vessel direction by Hessian analysis. Since the computation of flux was done by integrating gradient information at the sphere (vessel) boundary *at the original high resolution*, these methods claimed to be robust — unlike linear scale-space filters which involve blurring of detail — to situations in which vessels are adjacent to one another. This claim also applies to a filter recently proposed by Moreno and Smedby [195] who searched for ring-like patterns in the gradients at the surface of a sphere, exploiting properties of symmetry using measures of structuredness, evenness and uniformness.

Many authors have mixed linear and non-linear filters. Zana and Klein [196] followed the Laplacian with a sum of top-hat transforms. Mendonca *et al.* [197] followed difference-of-Gaussian filters with a sum of top-hat filters. Dufour [198] followed Hessian analysis with spatially variant morphological closing, using linear SEs aligned according to the local orientations obtained by the Hessian.

The model of a vessel as a tubular structure does not take into consideration

vessel branch points, which go undetected by many of these derivative filters. Recent filters have attempted to better model vessel junctions. Qian *et al.* [199] used a polar coordinate system to plot intensity profiles over the surface of sphere, and showed that at branch points there are three points at which vessels exit the sphere, which manifest as three narrow (high) intensity bands. Truc *et al.* [200] used a directional filter bank [201] to decompose the scan into a number of directional images containing line-like features. The benefit of Truc's method is twofold. Firstly, noise in the directional images is reduced due to its omnidirectional nature. Secondly, computation of the Hessian improves since only vessels of similar direction are considered, and this aids detection of vessels close to junctions.

5.2.2.2 Extraction of the vascular topology

Perhaps the most reliable methods of vessel tracking are those based on shortest path graph-based algorithms such as Dijkstra [202] which track between a given start and end point ("two-point" vessel tracking) [203, 27, 204, 205, 206, 207]. An image (or volume) is treated as a fully-connected lattice of nodes, one node per voxel, and a cost is associated with each edge between voxels, based on some measure of vessel path likelihood. The simplest method is to threshold nodes into vessel and background (by raw image intensity, or by some more complex vesselness measure as describe above), and then use a binary cost function that assigns a low cost to vessel and a high cost — perhaps infinity — to background. However, the user has to place two points for each vessel, and the process of navigating through the scan volume can become tiresome.

Other systems attempt more ambitious tracking of a vessel tree (frequently the coronary tree), with or without a single-seed user input at the root of the tree. Methods have variously used matched filtering [208], contour analysis [209, 210, 211], region growing [212, 213, 214, 197, 215, 216, 217, 218, 219], min-cost flow problem solving [220] and statistical methods [221, 222, 223]. Current approaches have capitalised on the success of machine learning algorithms. For example, Schneider [19] used a Hough regression forest to estimate vessel centre points. Cherry [20] trained regression forests to learn how to segment, terminate and then prune the tree of the marginal artery.

Such tree tracking algorithms with minimal or no user interaction require assumptions to be made (or learnt) about the orientation and depth of the tree, the number of branching levels, the vessel diameter, and the rate of contrast

attenuation. Prior anatomical knowledge may aid with these assumptions. In [224] Yim *et al.* took as input parameters the minimum branch length and number of bifurcations. In [225], Passat *et al.* segmented the superior sagittal sinus (principal vein of the human brain), incorporating information about the path of the vein relative to other structures of the brain for the purpose of generating cross-sectional planes through the vein and for initialising the segmentation, and also knowledge of the homogeneity of the vessel to justify a rolling average slice approach to segmentation. Chillet and Cool [226, 227] demonstrated the construction of a vascular atlas by computing a distance map (DM, the distance is to the nearest vessel centre) and then taking the mean and variance of all registered DMs to give a spatial estimate of vascular density. Passat *et al.* [214] went on to apply an atlas to vascular extraction, using an atlas of the head for the purpose of segmenting brain vessels. This atlas is divided into several geographical regions, each of which has homogeneous properties according to vessel size and orientation, and is used to inform a region-growing algorithm.

Finally, we mention that the technique of graph matching has been applied to vessel labelling, a task to which it is aptly suited [209, 228].

5.2.3 Motivation for our approach

In chapters 2 and 3 we have developed a random forest classification algorithm for the detection of landmarks. This motivated the idea of automatically detecting seed points, and hence using two-point tracking techniques to map the standard arterial system. This could provide a significant time saving over manual placement of seed points. To track the whole arterial tree as we have defined it (see Figure 5.8), the user would have to place between 21 (root + vessel termini only) and 39 points (including vascular branch points). Further, landmark detection allows easy handling of situations in which the scan acquisition region contains only part (or parts) of the vessel tree, since we can deduce by the positively detected landmarks which vessel sub-trees are present.

We have a pre-existing two-point tracking algorithm at TMVS based on the A* star algorithm. However, since the vessel being tracked is always known, we propose to augment this algorithm with anatomical information to aid the tracking process.

The ordering of the chapter is as follows. We start with a case study of landmark detection in cardiac CTA datasets, containing a mix of vascular and non-vascular landmarks, showing how feature parameters may be optimised for

the detection of landmarks in contrasted scans. We then define a set of vascular landmarks. We demonstrate how contextual information may be used to aid tracking, for the case of manually placed (ground truth) landmarks, presenting as part of this a comparison of the *non-linear* top-hat transform and the *linear* Frangi vesselness filter. Finally, we demonstrate the fully automatic vessel tracking system, tracking from detected landmarks.

An overview diagram of the vessel tracking system is shown in Figure 5.1.

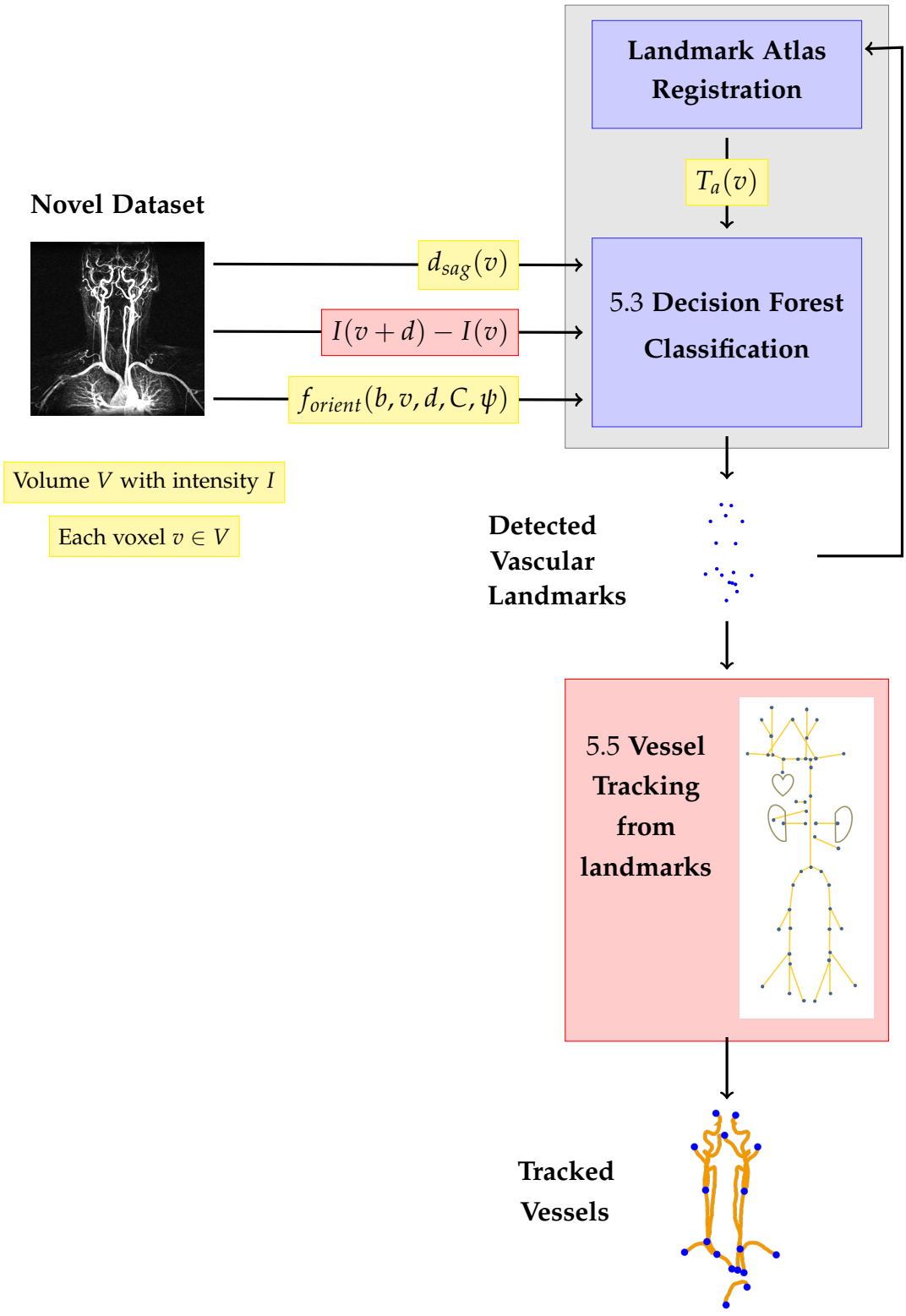
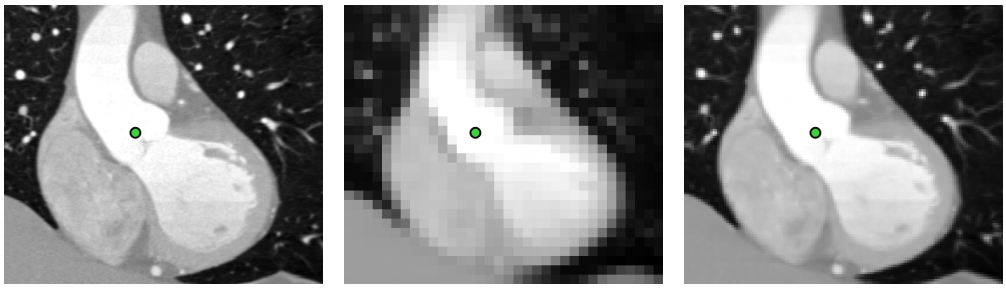


Figure 5.1: Overview diagram of the vessel tracking algorithm. Sagittal displacement $d_{sag}(v)$, relative intensity $I(v+d) - I(v)$, gradient orientation $f_{orient}(b, v, d, C, \psi)$ and atlas location $T_a(v)$ features are used for machine learning. In this chapter, we focus on tuning the intensity features for vascular landmarks, and the vessel tracking algorithm (components marked in pink).

(a) Right Coronary Ostium (b) 4mm voxel^{-1} (c) 1mm voxel^{-1}



(d) Aortic Valve (e) 4mm voxel^{-1} (f) 1mm voxel^{-1}

Figure 5.2: Images of two cardiac landmarks which are only visible at high resolution. Above: MIP images of the right coronary ostium a) at full resolution b) at 4mm voxel^{-1} resolution c) at 1mm voxel^{-1} resolution. Below: Slices through the aortic valve d) at full resolution e) at 4mm voxel^{-1} resolution f) at 1mm voxel^{-1} resolution. [TMVS Dataset IDs: 1437, 1461]

5.3 Vascular landmark detection

In this section, we take a look at improving the detection of vascular landmarks by moving to a higher resolution. This is motivated by the observation that many structures which we are attempting to landmark, in particular vascular landmarks, are not visible at 4mm voxel^{-1} resolution. This was highlighted in the discussion of chapter 2 (see Figure 2.26) and again in Figure 5.2.

We will show in section 5.3.1 that better accuracy can be achieved at higher resolution and with better feature parameterisation. In the light of a changing volume resolution, we then investigate the detection grid search interval d_{skip} in section 5.3.2 and show that there is a trade-off between accuracy and processing time.

5.3.1 Training a vascular landmark detector

We start by examining a 1mm voxel^{-1} detector trained on the cardiac subset of the datasets, to demonstrate the effect of different parameter combinations. For

each combination of parameters, we report the mean detection error, the AUC and the mean detection time.

5.3.1.1 Data

The data is comprised of the 45 cardiac datasets from the training cohort described in chapter 2. During testing, an out-of-bag strategy is employed, implemented by running detection using only those trees not trained on the dataset under test (see section 2.3.4).

5.3.1.2 Random forest detector

A random forest detector is trained using $T = 80$, $D_T = 15$ (from $D = 45$), $F_T = 2500$, $d_{max} = 52$mm (at 4mm voxel^{-1}) or 50mm (at 1mm voxel^{-1}), $w_{Node\_min} = 5.0$, $w_{Node\_Split\_min} = 2.0$ and $d_{skip} = 2$ voxels.

(See section 2.1 for symbol definitions.)

5.3.1.3 Sampling regions

The same total number of training samples are selected for the 1mm voxel^{-1} detector as for the original 4mm voxel^{-1}. This is achieved by shrinking $\sigma_{Sampling}$ by a quarter from 3.0mm to 0.75mm. As before, we use a ratio $B_{Ratio} = 5.0$ of background to landmark samples.

5.3.1.4 Landmarks

In this detector, seven landmarks are learnt. These are 1. *Bifurcation of trachea* 2. *Right coronary ostium* 3. *Left coronary ostium* 4. *Aortic valve (centre of the three semilunar cusps)* 5. *Heart apex (extremus in sagittal plane) at epicardium* 6. *Heart apex (extremus in sagittal plane) at endocardium* 7. *Left dome of diaphragm*.

5.3.1.5 Parameters for experimentation

Experiments were performed for all combinations of the following parameter values. A brief theoretical justification is given for why varying these parameters could give improved accuracy.

- **Resolution {4mm voxel^{-1}, 1mm voxel^{-1}}:** A 1mm voxel^{-1} resolution should allow the coronary arteries and valve cusps to be visible.

- **Voxel intensity values {Absolute, Relative}:** We propose that using relative intensity values $I(v+d) - I(v)$ (i.e. the intensity at an offset d from the voxel v minus that at v itself) will work better than absolute intensity values $I(v+d)$ for vascular landmarks, since the intensity of the contrast is variable across different scans. Relative landmark intensities may not work so well for landmarks on other tissues.

- **Feature size {Single voxel, Cuboidal $\leq$ 30mm, 50-50 mix}:** Cuboidal features refer to the mean intensity computed over all voxels within a *cuboid* C centred at $v+d$, as opposed to just the intensity at v. The dimensions are uniformly and randomly selected up to a maximum C_{max} = 30mm (as is done with gradient orientation features, see chapter 3). In theory, cuboidal features should provide information at multiple (larger) scales. Cuboidal features also make sense with higher resolution data due to the noise reduction from averaging across multiple voxel intensities. Additionally we try a 50-50 mix of the two as we did with some success for brain segmentation, see classifier B in chapter 4 i.e. 50% of trees trained on single intensity and 50% on cuboidal features.

- **Feature sampling pattern {Volumetric, Radial}:** *Radial* sampling refers to uniform sampling with respect to offset magnitude. *Volumetric* sampling refers to uniform sampling with respect to volume. See section 3.4.3 for illustration. If features are sampled more densely close to the landmark, as in *radial* sampling, then local appearance will have more weight, which should aid in our goal of precision.

5.3.1.6 Results

Table 5.1 shows the results. Relative intensities and radial feature sampling give significant improvement at both scales, making detectors 10 and 22 the best at resolutions D_{Res} of 4mm voxel^{-1} and 1mm voxel^{-1} respectively. Cuboidal features give no significant benefit. In terms of mean error, the high-resolution detector 22 performs the best, however it performs less well in terms of AUC because the trachea bifurcation is sometimes falsely detected when not present.

There is further a significant time penalty for the 1mm voxel^{-1} detector compared to the original 4mm voxel^{-1} resolution detector.

No.	Rank	D_{Res} mm voxel^{-1}	Intensity	Sampling	Box Size mm	Error mm	AUC	Time
1	22	4	Absolute	Volumetric	4	9.8	0.921	6s
2	19	4	Absolute	Volumetric	≤ 30	9.5	0.923	6s
3	13	4	Absolute	Volumetric	4, ≤ 30	8.9	0.928	6s
4	20	4	Absolute	Radial	4	9.6	0.923	4s
5	17	4	Absolute	Radial	≤ 30	9.3	0.946	4s
6	21	4	Absolute	Radial	4, ≤ 30	9.7	0.948	4s
7	16	4	Relative	Volumetric	4	9.2	0.963	5s
8	23	4	Relative	Volumetric	≤ 30	9.8	0.958	5s
9	18	4	Relative	Volumetric	4, ≤ 30	9.4	0.965	5s
10*	8	4	Relative	Radial	4	8.2	0.973	4s
11	15	4	Relative	Radial	≤ 30	9.2	0.951	3s
12	7	4	Relative	Radial	4, ≤ 30	8.2	0.975	3s
13	6	1	Absolute	Volumetric	1	8.0	0.963	3m
14	24	1	Absolute	Volumetric	≤ 30	10.1	0.934	3m
15	14	1	Absolute	Volumetric	1, ≤ 30	9.1	0.949	3m
16	4	1	Absolute	Radial	1	7.7	0.955	2m
17	11	1	Absolute	Radial	≤ 30	8.6	0.941	2m
18	10	1	Absolute	Radial	1, ≤ 30	8.5	0.953	2m
19	5	1	Relative	Volumetric	1	7.9	0.952	5m
20	12	1	Relative	Volumetric	≤ 30	8.9	0.944	5m
21	9	1	Relative	Volumetric	1, ≤ 30	8.2	0.941	5m
22*	1	1	Relative	Radial	1	6.4	0.892	2.5m
23	3	1	Relative	Radial	≤ 30	7.0	0.870	2.5m
24	2	1	Relative	Radial	1, ≤ 30	6.5	0.901	2.5m

Table 5.1: Parameter tuning for vascular landmarks in CT: Mean errors and run times. In the detection time column, *s* represents seconds and *m* represents minutes.

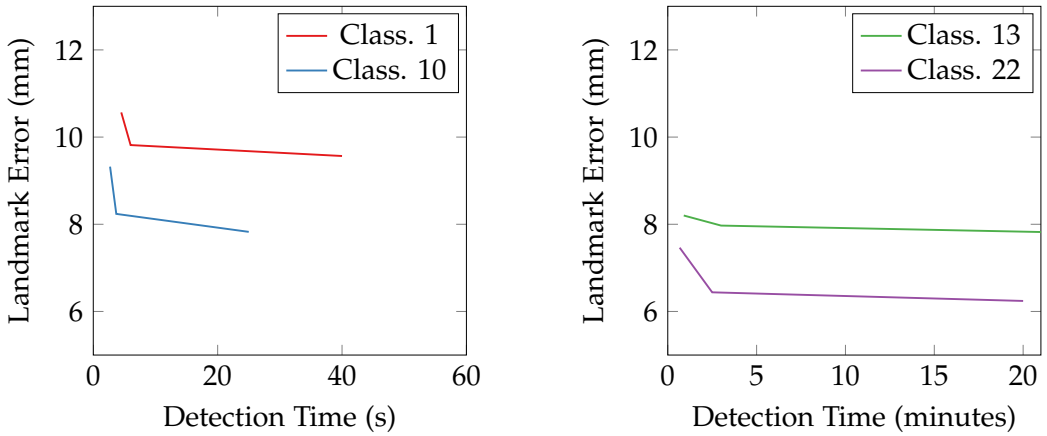
5.3.2 Trade-off between search density and accuracy

For efficiency, in all experiments so far, an initial global grid search has been run on the volume at intervals of $d_{skip} = 2$, in other words skipping every second voxel in each dimension, followed by a local search of the unseen voxels adjacent to the maximum probability voxel for each landmark (see section 2.3.2 for details). Using this strategy, the initial search is approximately eight times faster.

We have run a brief analysis of the original and optimised detectors at each resolution, to check the effect of d_{skip} on performance. Table 5.2 and Figure 5.3 show the results. It can be seen that a skip factor of two gives the lion's share of the performance, both in terms of accuracy and speed, for all detectors, validating this as a good compromise choice.

No.	Details	**Skip = 3**		**Skip = 2**		**No Skip**	
		Error	Time	Error	Time	Error	Time
1	$D_{Res} = 4$, Absolute, Uniform	10.6	4.5s	9.8	5.6s	9.6	40s
10	$D_{Res} = 4$, Relative, Radial	9.3	2.7s	8.2	3.7s	7.8	25s
13	$D_{Res} = 1$, Absolute, Uniform	8.2	54s	8.0	3m	7.8	22.5m
22	$D_{Res} = 1$, Relative, Radial	7.5	41s	6.4	2.5m	6.2	20m

Table 5.2: Results for different values of skip factor d_{skip}, measured in voxels. The detectors (Class. 1 etc.) are referred to by the numbers in Table 5.1. Errors are measured in millimetres. Time is measured in seconds (s) or minutes (m).



(a) 4mm voxel^{-1} detectors

(b) 1mm voxel^{-1} detectors

Figure 5.3: Graphs showing the trade-off between accuracy and detection time, as a function of d_{skip}. The detectors (Class. 1 etc.) are referred to by the numbers in Table 5.1.

5.3.3 Analysis of individual landmark errors

Figure 5.4 shows the per-landmark mean errors (both detector and inter-observer) for the 30 datasets where ground truth was collected by two observers.

The high resolution detector gives much improved performance in some cases (the bifurcation of the trachea and the dome of the diaphragm in particular) but debatable or worse performance in others (heart apex at endocardium). The possibility of combining different resolution results in future is discussed in section 5.8.2.

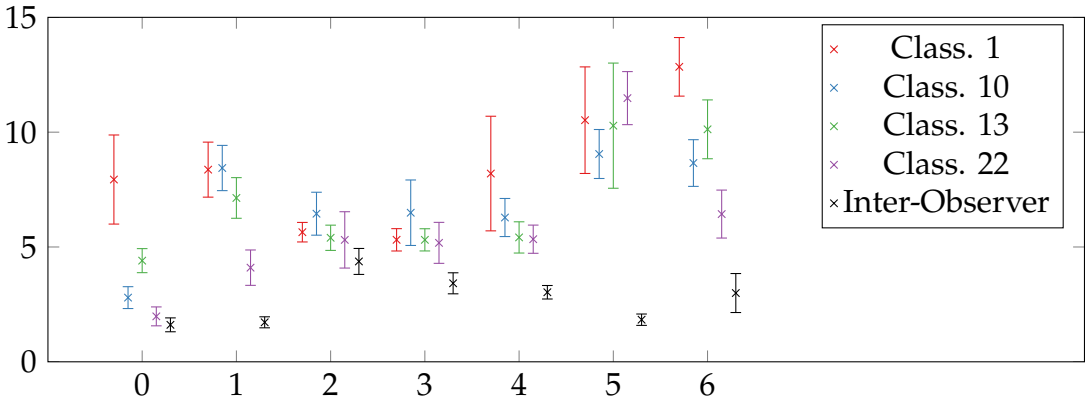
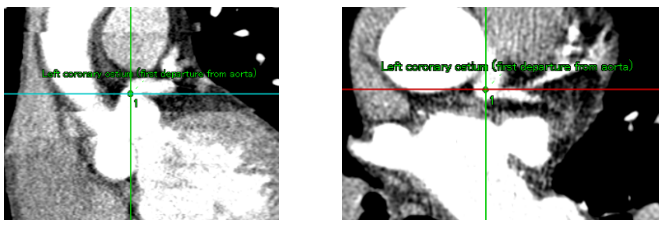


Figure 5.4: Graph showing error bars (mean error +/- standard deviation) for individual landmarks. 0 = *Bifurcation of trachea*, 1 = *Right Coronary Ostium*, 2 = *Left Coronary Ostium*, 3 = *Aortic Valve*, 4 = *Heart Apex in sagittal plane at epicardium*, 5 = *Heart Apex in sagittal plane at endocardium*, 6 = *Left dome of diaphragm*. The detectors (Class. 1 etc.) are named according to Table 5.1. The inter-observer errors are also given.

As a sidenote, it appears that there is significant difference between the inter-observer accuracies for the right and left coronary ostia. The left coronary artery has a much greater diameter at its origin which will make centring a point within the vessel more difficult. Further, the prescribed marking plane was not always used — Figure 5.5 illustrates the difference this can make.



(a) Coronal Plane (b) Axial Plane

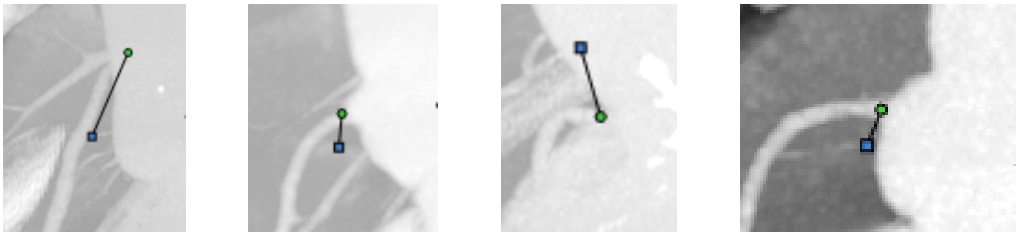
Figure 5.5: The choice of marking plane changes the apparent left coronary ostium centre location. [TMVS Dataset ID: 1465]

5.3.4 Visualising the probability clouds for selected landmarks

We now analyse the results by inspection of the images and of the landmark probability clouds.

Starting with vascular landmarks, we look at the origin of the right coronary artery, a landmark that was identified as problematic in chapter 2. Figure 5.6 shows four examples comparing the low resolution detector 10 with the high resolution detector 22. At high resolution (detector 22), the landmark is always located somewhere along the vessel, if not always at the origin. It would be possible to attempt vessel tracking using our tracking method (see section 5.5.5) from these results, even if they are imprecisely located.

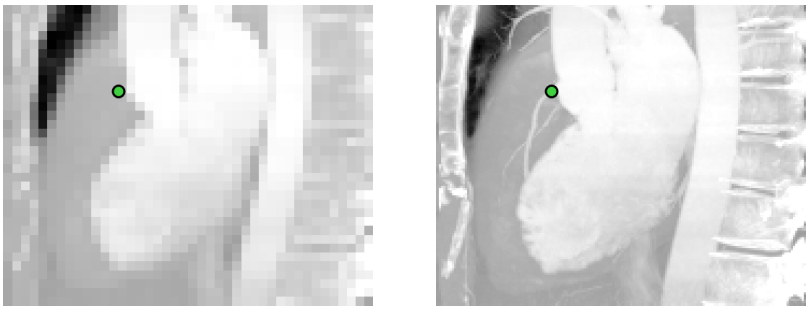
Figure 5.7 shows probability cloud images for dataset C. The shape of the probability cloud changes dramatically in response to the use of relative feature and radial sampling, and this explains why the detected landmark is always located somewhere along the path of the vessel.



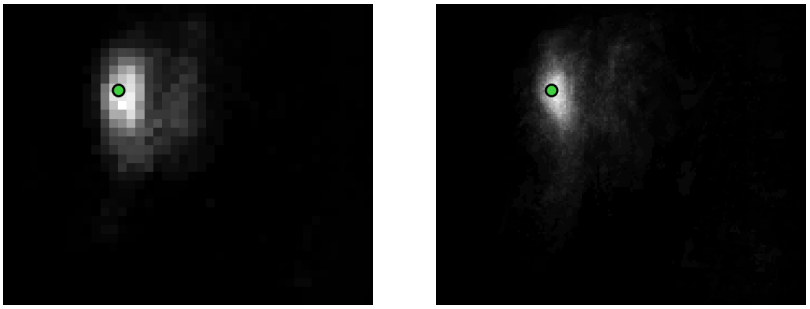
(a) A (Low) (b) B (Low) (c) C (Low) (d) D (Low)

(e) A (High) (f) B (High) (g) C (High) (h) D (High)

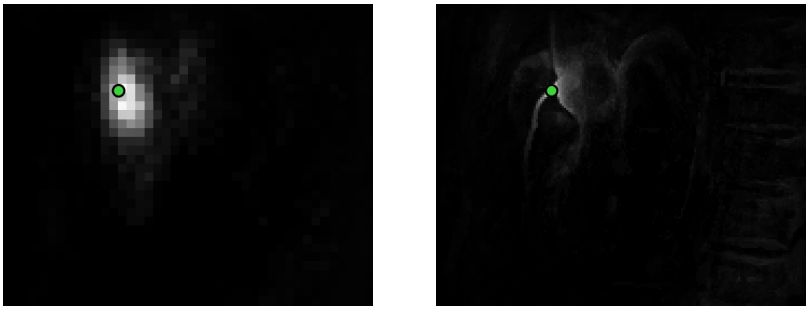
Figure 5.6: MIP images of the right coronary ostium in a few datasets, at 'Low' resolution (4mm voxel^{-1}) and 'High' resolution (1mm voxel^{-1}). In a) – d) the results of the optimised 4mm voxel^{-1} detector 10 are shown. In e) – h) the results of the optimised 1mm voxel^{-1} detector 22 are shown. At the higher resolution, the landmark is always located somewhere along the vessel, if not always at the origin. [TMVS Dataset IDs: 1435, 1437, 1442, 1447]



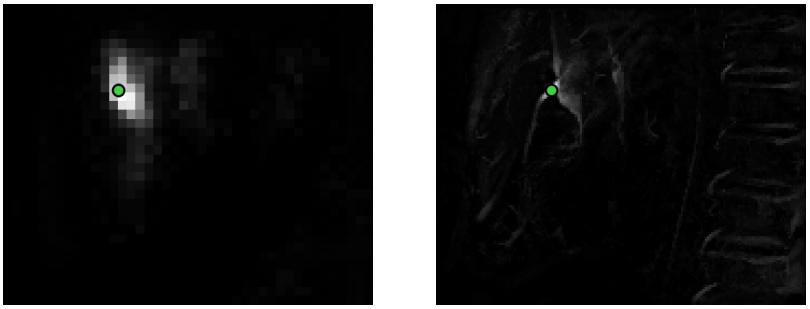
(a) 4mm Resolution (b) 1mm Resolution



(c) Detector 1 (d) Detector 13



(e) Detector 7 (f) Detector 19



(g) Detector 10 (h) Detector 22

Figure 5.7: MIP images and probability clouds for the right coronary ostium for dataset B of Figure 5.6, showing the effect of resolution, sampling strategy and relative intensities. The left and right columns show 4mm voxel^{-1} and 1mm voxel^{-1} resolution images respectively. Detectors 1 and 13 use the original parameters, detectors 7 and 19 use relative intensities, and detectors 10 and 22 use both relative intensities and radial sampling. Intensities are scaled linearly from black (minimum probability present) to white (maximum probability present). [TMVS Dataset ID: 1437]

5.4 Defining arterial tree landmarks

For the purpose of vessel tracking, a new set of landmarks specific to the vascular system were defined. A graphical representation of the standard arterial tree is shown in Figure 5.8. Major arteries are represented as arcs connecting the vessel branch points and terminus nodes. The yellow lines in the schematic indicate vessels in the standard arterial system. A full list of vessels and vascular landmarks is given in appendix B.

According to this scheme, each artery may be uniquely expressed by a pair of landmark points at its proximal and distal ends. The vessel tree is expressed as the union of a number of vessel segments, for each of which the origin and terminal landmarks are known.

There are a number of studies in the literature surveying vascular variation in different parts of the body [229, 230, 231]. Indeed, our cohort of 53 patients contained several instances of variation, including: a bovine arch, a left vertebral artery arising from the aorta, accessory renal arteries, and different branching configurations in the lower legs. For the work in this chapter, variant arteries are ignored. Dealing with variation is a subject for future work. A landmarking approach may not be the way to go since many landmarks would be required for adequate representation.

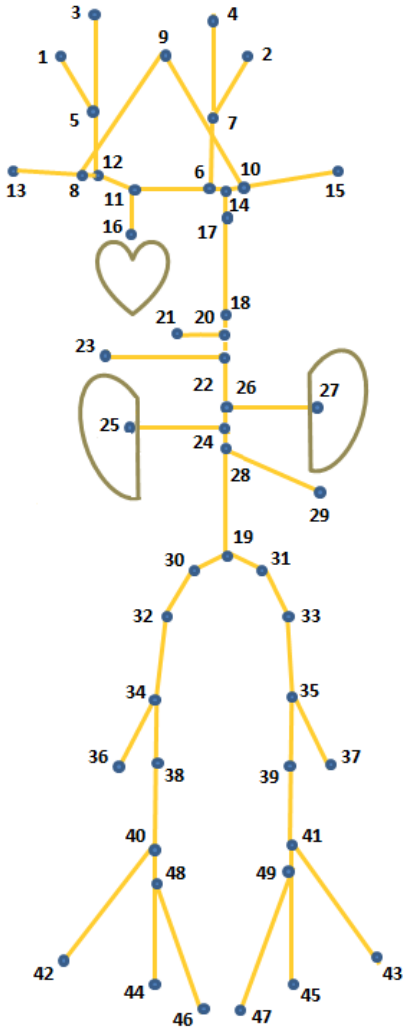


Figure 5.8: Schematic of the vascular landmarks and vessels in the standard arterial system. The blue dots represent landmarks, and the yellow lines represent arteries. A key to the numbered landmarks and a full list of vessels is given in appendix B.

5.5 Vessel tracking from manually placed landmarks

In this section we examine how contextual information can improve semi-automated vessel tracking. We begin with a pre-existing algorithm for tracking between two manually placed points. This algorithm works well for large, bright, non-tortuous vessels, but gives less reliable results for more complicated cases (see results for the "control" method in Table 5.3).

Since we have pre-defined the seed points in section 5.4, we have knowledge of which vessels are being tracked. From this, we can make predictions about vessel properties such as diameter, tortuosity and shape, all of which are useful to guide tracking. Sections 5.5.3 and 5.5.4 describe how we exploit knowledge about the typical vessel diameter and vessel path in order to improve the tracking results.

5.5.1 Image acquisition

Fourteen Contrast Enhanced MRA scans were obtained from symptomatic subjects with suspected arterial pathology. Twelve scans come from a Siemens scanner and are study stitched whole-body scans (excluding the upper arms) for which the component data comes from four overlapping stations: head and neck, thorax, abdomen and upper part of lower limbs, lower part of lower limbs. Two scans come from Toshiba scanners; one is of the head and neck, and one is of the abdomen.

5.5.2 Ground truth collection

For each of the 14 datasets, an anatomical expert:

- Marked points for all vascular landmarks that are present in the scan (from the set defined in section 5.4).
- Manually plotted a centre line for each vessel that is present in the scan.

The ground truth landmarks are used to seed the tracking algorithm (see section 5.6 for results from automatically detected landmarks). We use manually placed landmarks in the first place, in order to evaluate the tracking algorithm independently of the accuracy of the detection algorithm.

The centre lines are used to create the vascular atlases (see section 5.5.4) and to compute evaluation metrics.

5.5.3 Introducing a prior for vessel size: Applying a vesselness filter

Two filters were considered, the white top-hat transform and the Frangi vesselness filter.

5.5.3.1 White top-hat transform

The white top-hat transform [170] is a morphological operation comprising grey-scale erosion, followed by grey-scale dilation of equal magnitude, and finally subtraction from the original image.

We choose to use a "flat" 2D square structuring element E with sides of length $2r$, applied in each of the three planes of the volume, ψ={*axial, coronal, sagittal*}. As was done by Murphy [232], 2D operators are applied rather than a single cuboidal 3D operator in order to avoid preserving sheet structures.

In this case, for a volume with intensities $I(v)$, grey-scale erosion $\ominus$ simply consists of running through all voxels in the local neighbourhood and assigning the minimum value as in equation 5.1. So, for a plane ψ where x and y represent the two dimensions of the plane:

$$(I \ominus E)(v) = \min_{x=-r}^{r}[\min_{y=-r}^{r}[I(v + (x, y))]] \tag{5.1}$$

Grey-scale dilation $\oplus$ then consists of assigning the *maximum* value to the resulting volume, as shown in equation 5.2.

$$(I \oplus E)(v) = \max_{x=-r}^{r}[\max_{y=-r}^{r}[(I \ominus E)(v + (x, y))]] \tag{5.2}$$

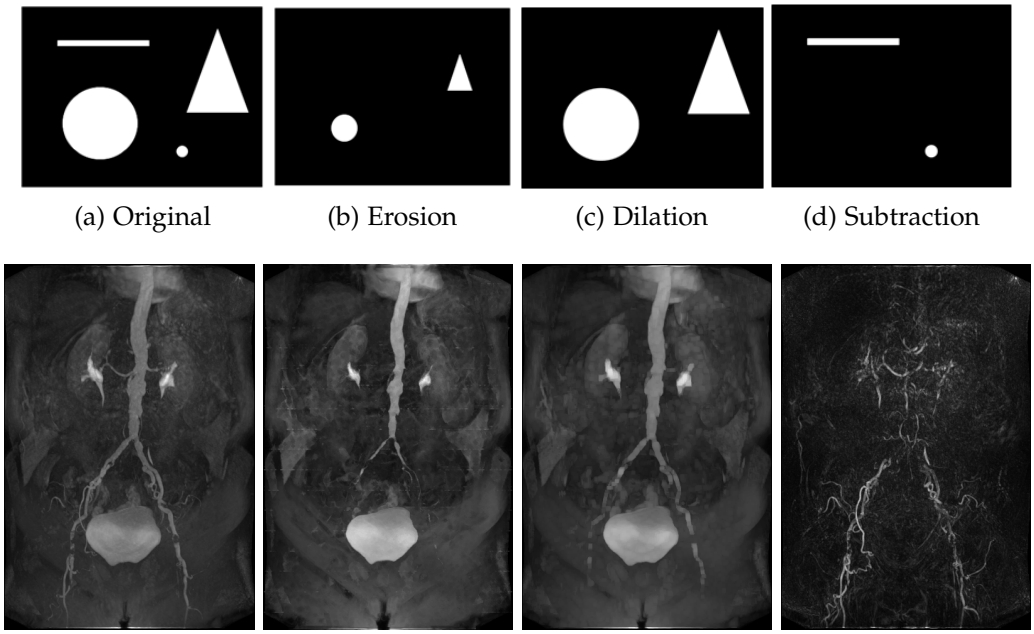
Finally, the maximum of the three plane results is found.

$$I_{3-way}(v) = \max_{\psi}((I \ominus E) \oplus E) \tag{5.3}$$

Equation 5.4 shows the full top-hat operation.

$$F_{Top\_Hat}(v) = I(v) - I_{3-way}(v) \tag{5.4}$$

The white-top hat transform should enhance vessels — and other small bright structures — with radius smaller than or equal to r, and suppress large background structures. This is illustrated in Figure 5.9. We try a range of sizes equivalent to the range of vessel diameters in which we are interested. Figure 5.17 shows results for real data for two different filter sizes.



(a) Original (b) Erosion (c) Dilation (d) Subtraction

(e) Original (f) Erosion (g) Dilation (h) Subtraction

Figure 5.9: Images showing the stages of application of the top-hat transform. Top: Binary image. Bottom: MIP images from a real (grey-scale) dataset in which a filter at scale r=2mm is used. Erosion and dilation operations are performed in sequence, and the result is subtracted from the original image to leave only the small bright structures visible. [TMVS Dataset ID: 3738]

5.5.3.2 Vesselness filter

This is a filter by Frangi *et al.* [182], based on analysing the eigenvalues of the Hessian matrix at a given scale and location. According to scale-space theory [188], the derivatives of a image $L(v)$ at the scale s may be found by convolving $L(v)$ with the corresponding partial derivatives of the zeroth-order Gaussian $G(v,s)$. So, for a second derivative:

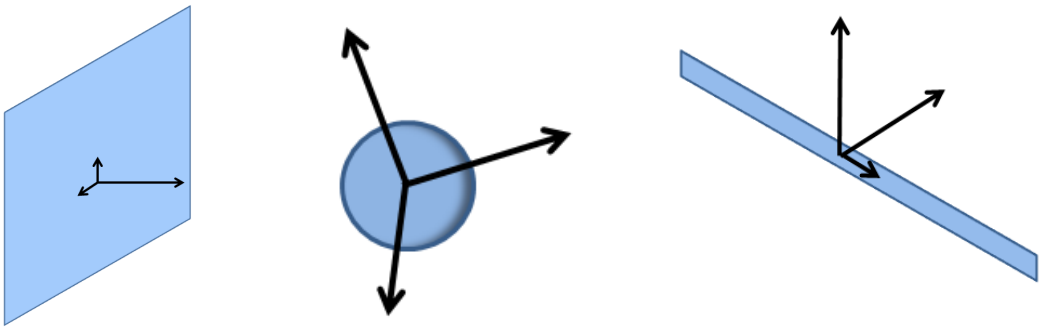
$$\frac{\partial^2}{\partial v^2} L(v,s) = L(v) * \frac{\partial^2}{\partial v^2} G(v,s) \tag{5.5}$$

We re-arrange the ordering of the operations. This gives the same result since the (discrete) Gaussian smoothing operator commutes with the (discrete) second-order central difference operator that we employ [189].

$$\frac{\partial^2}{\partial v^2} L(v,s) = \frac{\partial^2}{\partial v^2} \left(L(v) * G(v,s) \right) \tag{5.6}$$

In order that the results are comparable across scales, we add a normalisation factor (s^2), as shown in equation 5.7.

$$\frac{\partial^2}{\partial v^2} L(v,s) = s^2 \frac{\partial^2}{\partial v^2} \left(L(v) * G(v,s) \right) \tag{5.7}$$



(a) Sheet (b) Blob (c) Cylinder

Figure 5.10: Illustration of the three Hessian matrix eigenvectors for sheets, blobs and cylinders, indicated with black arrows. a) Sheet: The eigenvector perpendicular to the sheet is large, whereas those in the plane are small. b) Blob: The eigenvectors are large in all directions. In fact, for a perfectly spherical "blob", the orientation of the eigenvectors is arbitrary. c) Cylinder/Vessel: The eigenvector in the direction of the vessel is small. Those aligned with the cylinder cross-section are large.

The Hessian matrix components (second derivatives) will be large in magnitude if computed at a peak in intensity. Second derivatives are *positive* in value at maxima i.e. bright structures, and *negative* in value at minima i.e. dark structures. Eigenvalue decomposition extracts three orthogonal directions, with the largest eigenvector oriented in the direction of the largest derivative. Different structure shapes exhibit different patterns, as shown in Figure 5.10. For a vessel, the smallest eigenvector corresponds to the direction of the vessel, with the two

larger eigenvectors oriented perpendicular to it. This can be distinguished from blobs where all vectors are large, and sheets where only one vector is large. The Frangi vesselness filter equation 5.11 contains terms, R_A and R_B, which control the preference for thin vessel structures over sheets or blobs. In addition there is a noise-suppression term S, which is the Frobenius matrix norm of the Hessian. Given the eigenvalues λ_1, λ_2 and λ_3, numbered by magnitude from smallest to largest, we describe the terms of the Frangi filter below.

$$R_A = \frac{(\text{Largest Cross Section Area})/\pi}{(\text{Largest Axis Semi-Length})^2} = \frac{|\lambda_2|}{|\lambda_3|} \tag{5.8}$$

$$R_B = \frac{\text{Volume}/(4\pi/3)}{(\text{Largest Cross Section Area}/\pi)^{3/2}} = \frac{|\lambda_1|}{\sqrt{|\lambda_2\lambda_3|}} \tag{5.9}$$

$$S = \sqrt{\sum_{k=1}^{3} \lambda_k^2} \tag{5.10}$$

$$F_{Frangi}(s) = \begin{cases} 0 & \text{if} \quad \lambda_2 > 0 \text{ or } \lambda_3 > 0 \\ (1 - exp\left(-\frac{R_A^2}{2\alpha^2}\right))exp\left(-\frac{R_B^2}{2\beta^2}\right)(1 - exp\left(-\frac{S^2}{2c^2}\right)) & \end{cases} \tag{5.11}$$

In our experiments, we set the parameter values to $\alpha = 0.5$, $\beta = 0.5$ and $c = 500$. Values were chosen by visual assessment of a few randomly chosen training datasets.

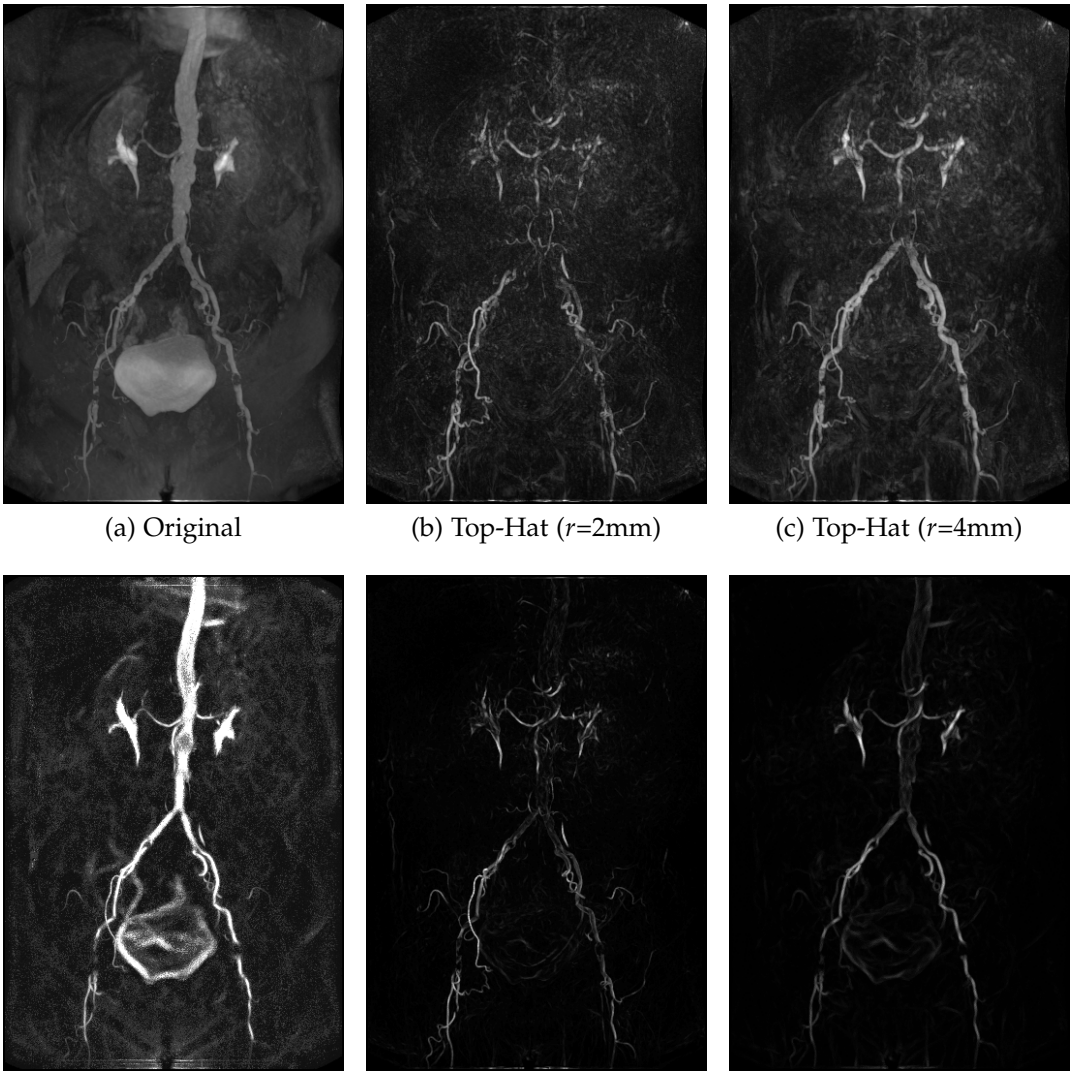
In theory, the scale at which the filter response peaks should correspond to the radius of the vessel being measured. In practice, we found that maximum responses were obtained at a rather smaller scale (see Figure 5.12).

5.5.3.3 Filter comparison

A visual comparison of the filters is given in Figure 5.11. We now attempt to evaluate the filters numerically.

Ideally, the vessel filter should have high sensitivity and specificity for voxels belonging to the vessel of interest. Sensitivity is particularly important, so as not to misclassify vessel voxels as background.

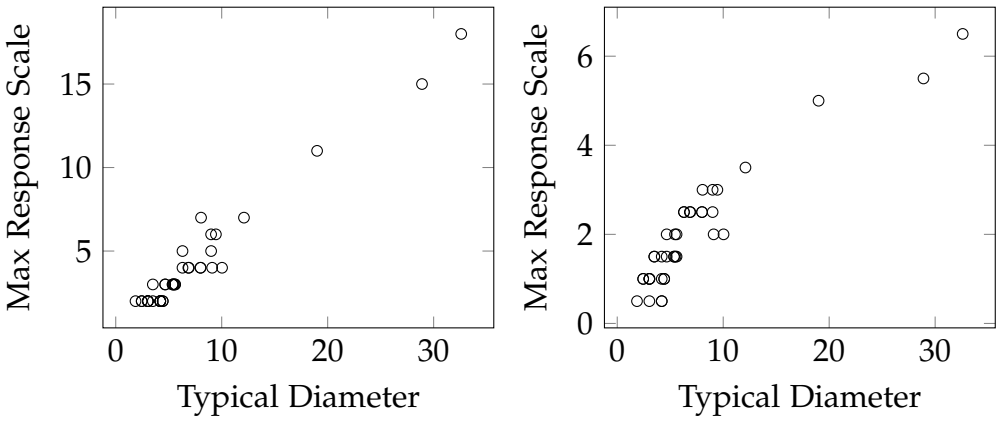
A comparison of the two filters was performed by computing for each vessel



(a) Original (b) Top-Hat (r=2mm) (c) Top-Hat (r=4mm)

(d) Frangi (all scales) (e) Frangi (s=1mm) (f) Frangi (s=2mm)

Figure 5.11: MIP images showing comparison of filtering using the top-hat transform and the Frangi filter. The same dataset is used as in Figure 5.9. a) Original image. d) Cumulative result for the Frangi vesselness filter. Two approximately equivalent scales are then shown for each filter. [TMVS Dataset ID: 3738]



(a) Top-hat transform

(b) Frangi vesselness filter

Figure 5.12: Graphs showing the correlation between typical vessel diameter and the filter size at which the maximum response is obtained (see Figure 5.13). Left: Top-hat transform. Right: Frangi vesselness filter. Typical vessel diameters are taken from Avolio [233] and Kahraman et al [234].

the filter responses $F(c)$ for vessel centre line voxels c over a range of filter sizes, and recording the 10th, 50th (median) and 90th percentile response values. Filter response is measured in units of statistical z-score, see equation 5.12.

$$R_{Filter} = \frac{F(c) - \mu}{\sigma} \tag{5.12}$$

Here μ and σ refer to the mean and standard deviation of the filter responses, computed over all voxels (vessel and non-vessel) in the training volumes.

Results are shown in Figures 5.12 and 5.13. There is a clear positive correlation between the predicted vessel diameter and the filter size at which the response peaks. The top-hat transform is most sensitive, particularly in the case of the small vessels of the lower leg, which are dim and may be just one or two voxels in diameter. On the other hand, the Frangi filter does not well distinguish vessel from noise — see the responses for the right posterior tibial artery; the bottom 10% of voxels are eliciting almost zero response. This problem is referred to in the paper on scale-space by Florack *et al.* [188] which states that the approximation to a scale s is only accurate where $s \gg s_0$, s_0 being the pixel scale. This is because at low resolution a Gaussian kernel has little meaning. A further problem is that the vesselness filter works poorly at branch points, as mentioned in the prior art section; this is evident in the images of Figure 5.11.

It was found empirically that the top-hat transform was indeed the most

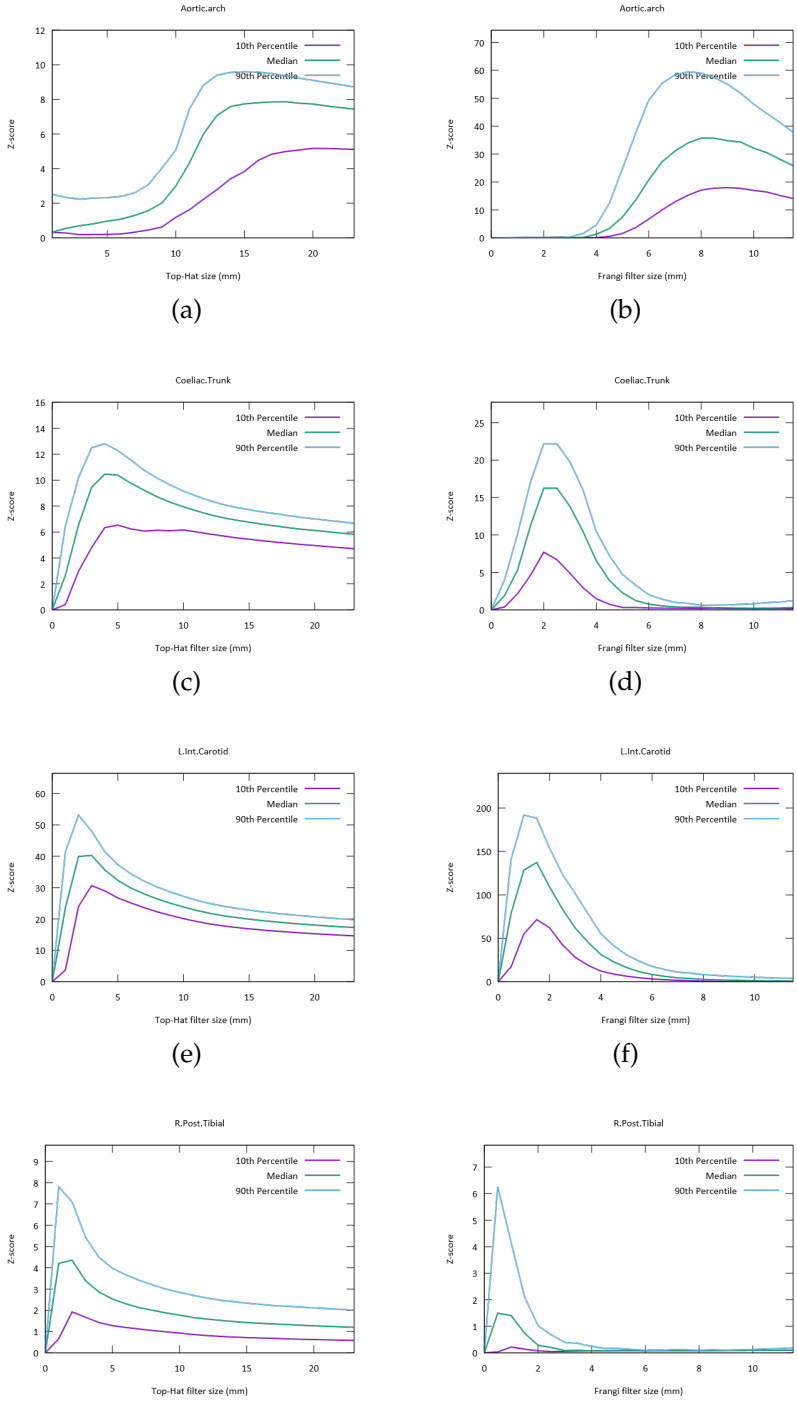


Figure 5.13: Graphs showing the filter responses for four different vessels, ordered by diameter from largest (top) to smallest (bottom): Aortic arch, coeliac trunk, internal carotid artery and right posterior tibial artery. Left: Top-hat transform. Right: Frangi vesselness filter. The horizontal axes show filter size i.e. Top-hat cuboid half-size r and Frangi scale s. The vertical axes show Z-score values. The 10th, 50th (median) and 90th percentile Z-scores are shown in blue, green and purple respectively.

effective for the purpose of tracking. We note that the top-hat transform is basically a low-pass filter. There is the option of combining two filters to give a band-pass filter (i.e. minus a smaller filter result from a larger filter result). By doing this, we would risk suppressing stenosed regions. Currently, we run the opposite risk, of filtering out aneurysms. This is an issue to be aware of when considering the vessel tracking application.

5.5.4 Introducing a prior for vessel path: Creating a vascular atlas

Vessels have characteristic shapes. If the vessel being tracked is known, then it is possible to predict the path of the vessel to some degree. We develop a simple method to create a "vascular atlas" for each vessel, which represents our prior belief (as per Bayesian reasoning) for the probability that an edge between two voxels belongs to the vessel of interest, given its position and direction. The prior probability doubles up as the atlas cost function C_π (see later how this is used in the tracking method).

The vascular atlas is constructed in a leave-one-out fashion. Non-rigid alignment between scans is found using a 3D thin plate spline [59], created from the detected vascular landmark locations. The registered ground truth centre lines are then resampled so that they contain data points at a spacing corresponding to the test scan resolution. Kernel density estimation is used to construct a continuous probability density function, as a function of the centre line ground truth points $t = 1 \ldots n$ with position $\mathbf{p}_t$. The contribution of each training point p_t to the cost at p is weighted by a Gaussian function $g_t = G(\mathbf{p}_t|\mathbf{p}, \sigma_\pi)$, parameterised by σ_π. Were we to simply construct a density map, then we take $\mathbf{p}$ to be the position of the destination voxel for the edge being costed and obtain:

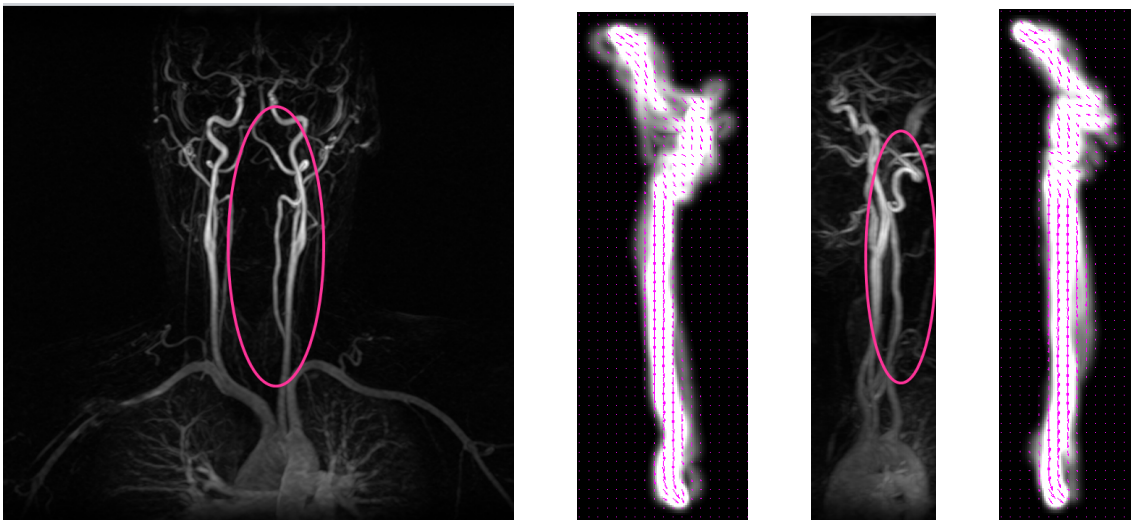
$$C_\pi(\mathbf{p}) = K_N \sum_{t=1}^{n} g_t \tag{5.13}$$

where the normalising factor K_N is selected such that the maximum value of the cost function C_π which is present in the atlas is one i.e. C_π ranges between zero and one.

However, we also take account of the *orientation* $\mathbf{o}$ of an edge, compared to those in the ground truth. Therefore we augment equation 5.13 with the orientations $\mathbf{o}_t$, of the edges connecting each point t with the previous point $t-1$.

$$C_\pi(\mathbf{p}, \mathbf{o}) = K_N \left(\sum_{t=1}^{n} g_t - \left| \sum_{t=1}^{n} g_t \hat{\mathbf{o}}_t \right| + \hat{\mathbf{o}} \cdot \sum_{t=1}^{n} g_t \hat{\mathbf{o}}_t \right) \tag{5.14}$$

An example map is shown in Figure 5.14 for the left vertebral artery.



(a) (b) (c) (d)

Figure 5.14: Coronal and sagittal MIP images showing the vessel flow probability map for the left vertebral artery. The image intensity represents the magnitude of the term $\sum_{t=1}^{n} g_t$, and the arrows show the direction and size of the vector $\sum_{t=1}^{n} g_t \hat{\mathbf{o}}_t$ at each point. Vector field is sampled at low resolution for ease of visualisation. a) Coronal view of the LVA. b) Coronal view of the LVA probability atlas c) Sagittal view of the LVA d) Sagittal view of the LVA probability atlas.

5.5.5 Tracking algorithm

A top-hat transform is used for filtering. For efficiency only, two top-hat filter sizes were used (2mm and 4mm half-size). For tracking each vessel, we select either the original volume or one of the filtered volumes according to the filter choice in Table 5.3. The best filter for each vessel was chosen based on which gave the best overall tracking results.

The A* algorithm [235] is used to efficiently search the space of the selected volume to find the shortest path between each pair of landmarks. This algorithm is a variant of Dijkstra's shortest path algorithm [202], which uses a priority queue to reduce the proportion of the space that is searched, by consideration of the Euclidean distance to the end node, or the *future path cost*. Each voxel in the

scan is treated as a node on the graph, and each voxel is considered to have an edge to connecting it to each of its 26 neighbouring nodes.

The cost of an edge is a combination of an intensity-based cost derived from the intensity I of the destination node, and a location-based cost computed using a vascular probability atlas constructed from ground truth.

Given a background intensity threshold I_B and a vessel intensity threshold I_V found using the convex hull method of histogram analysis [236], the cost due to intensity C_I is as follows.

$$C_I(I) = \begin{cases} k_I & \text{if} \quad I < I_B \\ 1 + (k_I - 1)\frac{(I_V - I)}{(I_V - I_B)} & \text{if} \quad I > I_B \text{ and } I < I_V \\ 1 & \text{if} \quad I > I_V \end{cases} \tag{5.15}$$

Misclassification may reasonably occur in cases of stenosis, in which there is a small region of non- or low contrast at the point of stenosis. For this reason, we allow tracking of the vessel through background, subject to a cost penalty. The constant k_I controls the relative preference for bright voxels ($I > I_V$) versus background vessels ($I < I_B$). A higher value is beneficial for vessels with high tortuosity, or where the vessel flow probability map is not considered to be a strong model.

The cost due to intensity C_I and the cost due to the vascular probability atlas $C_\pi(\mathbf{p}, \mathbf{o})$ are combined, and scaled by the length d of the edge. In a 3D square grid, d will be equal to 1, $\sqrt{2}$ or $\sqrt{3}$, depending on whether the edge travels in one, two or three dimensions between neighbouring nodes. Background voxels $C_\pi < t_\pi$ are not considered, where t_π is an empirically chosen threshold.

$$C(I, \mathbf{p}, \mathbf{o}) = \begin{cases} \infty & \text{if } I < I_B,\ C_\pi(\mathbf{p}, \mathbf{o}) \leq t_\pi \\ d\big(C_I(I)(1.0 + (k_\pi - 1)(1 - C_\pi(\mathbf{p}, \mathbf{o})))\big) & \text{otherwise} \end{cases} \tag{5.16}$$

The cost function $C(I, \mathbf{p}, \mathbf{o})$ is illustrated in Figure 5.15.

The resulting path will take a 'racing line' within the vessel. The path is refined to a centre line by performing vessel contour finding in 2D slices orthogonal to the path — not described here.

Fixed limits are set for the total cost of the path and the total number of voxels which are visited, such that in cases where there is no obvious solution, the algorithm fails rather than returning e.g. a direct path between vascular landmark

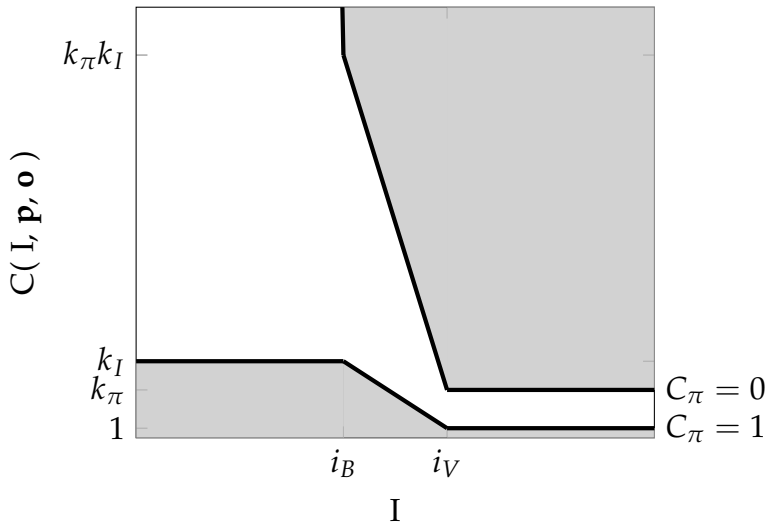


Figure 5.15: Graph showing the vessel tracking cost function $C(I, \mathbf{p}, \mathbf{o})$ (see equation 5.16). Raw voxel intensity is shown on the horizontal axis. Each voxel is assigned a cost according to its vascular probability map value C_π, from the range shown by the white region: the lower limit corresponds to an atlas cost of 1.0, the upper limit corresponds to an atlas cost of 0.0 and the cost transitions linearly in between. The grey shading indicates non-accessible cost regions.

end points which travels predominantly via background voxels.

5.5.6 Results

Parameter values of $k_I = 4$, $k_\pi = 2$ and $t_\pi = 0.6$ were chosen.

Results are shown in Table 5.3 for tracking vessels from manually placed landmarks. The control experiment and the context-aware method are compared. The control experiment is the original method without the addition of contextual knowledge i.e. no pre-processing filtering step and no atlas prior cost element (C_π is set to 1.0).

Evaluation Metric: The evaluation metric, % *Tr*, is the mean percentage of the vessel which is tracked successfully. This is calculated as follows:

- The ground truth centre line is matched to the generated centre line so that the sum Euclidean distance between them is minimal (see evaluation method described by Schaap [237]).
- The percentage of ground truth points is computed which have a corresponding generated point within the typical vessel radius.

Typical vessel diameters are taken from Avolio [233] and Kahraman et al [234].

Table 5.3: Vessel tracking results from manually placed landmarks in MRA datasets. n = number of test examples, S_D is the typical diameter according to Avolio [233] and Kahraman et al [234], S_F is the top-hat filter size = dimension of cubic filtering element, *Control* = results for control experiment, *Filter* = results using filtering only, *Atlas* = results using atlas only, *Both* = results for full method i.e. both filter and atlas. The results are expressed as the mean percentage of the vessel tracked successfully. The best result for each vessel is highlighted in grey.

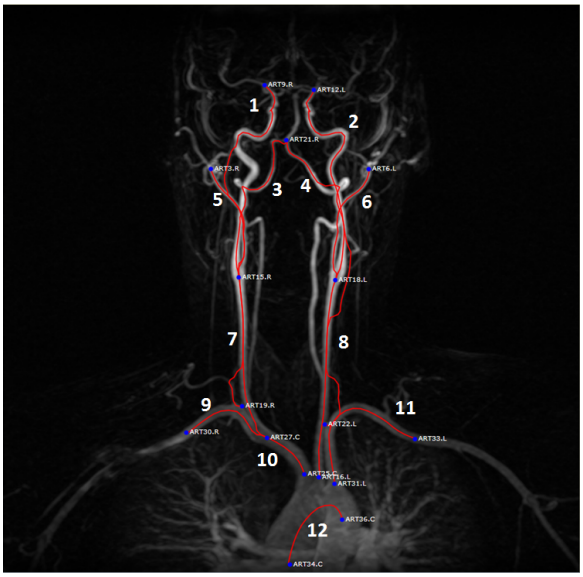
Artery	**n**	S_D **(mm)**	S_F **(mm)**	**Control** % *Tr*	**Filter** % *Tr*	**Atlas** % *Tr*	**Both** % *Tr*
R External Carotid	13	3.0	∞	69.1	69.1	71.2	71.2
L External Carotid	13	3.0	∞	89.1	89.1	88.5	88.5
R Internal Carotid	13	3.5	∞	79.0	79.0	83.2	83.2
L Internal Carotid	13	3.5	∞	85.0	85.0	87.3	87.3
R Common Carotid	13	6.9	∞	84.3	84.3	84.8	84.8
L Common Carotid	13	6.9	∞	87.6	87.6	89.2	89.2
R Vertebral	12	4.4	∞	69.5	69.5	74.0	74.0
L Vertebral	11	4.4	∞	83.4	83.4	87.9	87.9
Brachiocephalic	14	12	∞	100	100	100	100
R Subclavian	14	9.0	∞	94.4	94.4	99.9	99.9
L Subclavian	14	9.0	∞	100	100	100	100
Aortic Arch	14	33	∞	93.2	93.2	100	100
Thoracic Aorta	13	29	∞	100	100	100	100
Abdominal Aorta	12	19	∞	100	100	100	100
Coeliac Trunk	13	9.1	∞	92.3	92.3	99.5	99.5
Superior Mesenteric	13	10	9	53.8	85.2	61.5	93.1
Right Main Renal	13	5.5	9	84.6	100	100	100
Left Main Renal	12	5.5	9	82.2	94.7	85.7	98.9
Inferior Mesenteric	10	1.9	5	27.7	43.8	28.5	43.1
R Common Iliac	12	8.1	∞	91.4	91.0	91.0	91.0
L Common Iliac	12	9.5	∞	91.0	91.0	91.3	91.3
R External Iliac	12	8.0	∞	91.7	91.7	91.7	91.7
L External Iliac	12	8.0	∞	100	100	100	100
R Common Femoral	12	6.3	∞	100	100	100	100
L Common Femoral	12	6.3	∞	100	100	100	100
R Profunda Femoris	12	5.4	9	94.0	99.7	96.2	98.6

Continued on next page

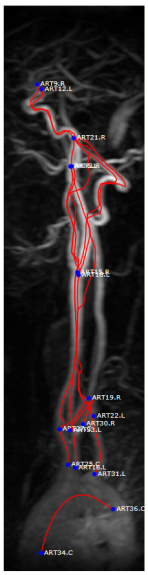
Artery	**n**	S_D **(mm)**	S_F **(mm)**	**Control** % *Tr*	**Filter** % *Tr*	**Atlas** % *Tr*	**Both** % *Tr*
L Profunda Femoris	12	5.4	9	91.0	96.1	92.7	97.3
R Superior Femoral	11	5.6	9	9.1	72.7	9.1	72.7
L Superior Femoral	12	5.6	9	56.0	75.0	58.0	75.0
R Popliteal	10	4.7	9	79.9	100	90.0	100
L Popliteal	12	4.7	9	83.3	91.7	89.0	96.1
R Anterior Tibial	8	2.5	5	22.9	72.9	35.0	73.2
L Anterior Tibial	9	2.5	5	34.0	93.4	82.4	92.4
R Tibioperoneal Trunk	11	4.2	5	98.8	98.4	99.2	98.3
L Tibioperoneal Trunk	11	4.2	5	89.0	100	98.3	100
R Peroneal	8	3.0	5	59.2	80.5	60.1	80.7
L Peroneal	9	3.0	5	52.6	86.2	52.4	85.7
R Posterior Tibial	11	4.2	5	43.2	69.7	44.2	75.2
L Posterior Tibial	9	4.2	5	24.2	72.6	35.2	72.6
Overall	460			78.8	88.6	82.3	90.3

It can be seen that for smaller vessels, the use of contextual information increases tracking success. For larger vessels, the control and the full method perform similarly well. Small dim vessels such as in the lower legs appear to benefit most from vessel enhancement by filtering. Small brighter vessels in regions with many other vessels, such the head, appear to most benefit from the atlas prior on the likely path of the vessel.

One-tail paired *t*-tests were performed for the overall figures for percentage vessel tracked, giving *p*-values less than 0.001, substantiating that there is a statistically significant overall improvement to vessel tracking.



(a) Control, Coronal View

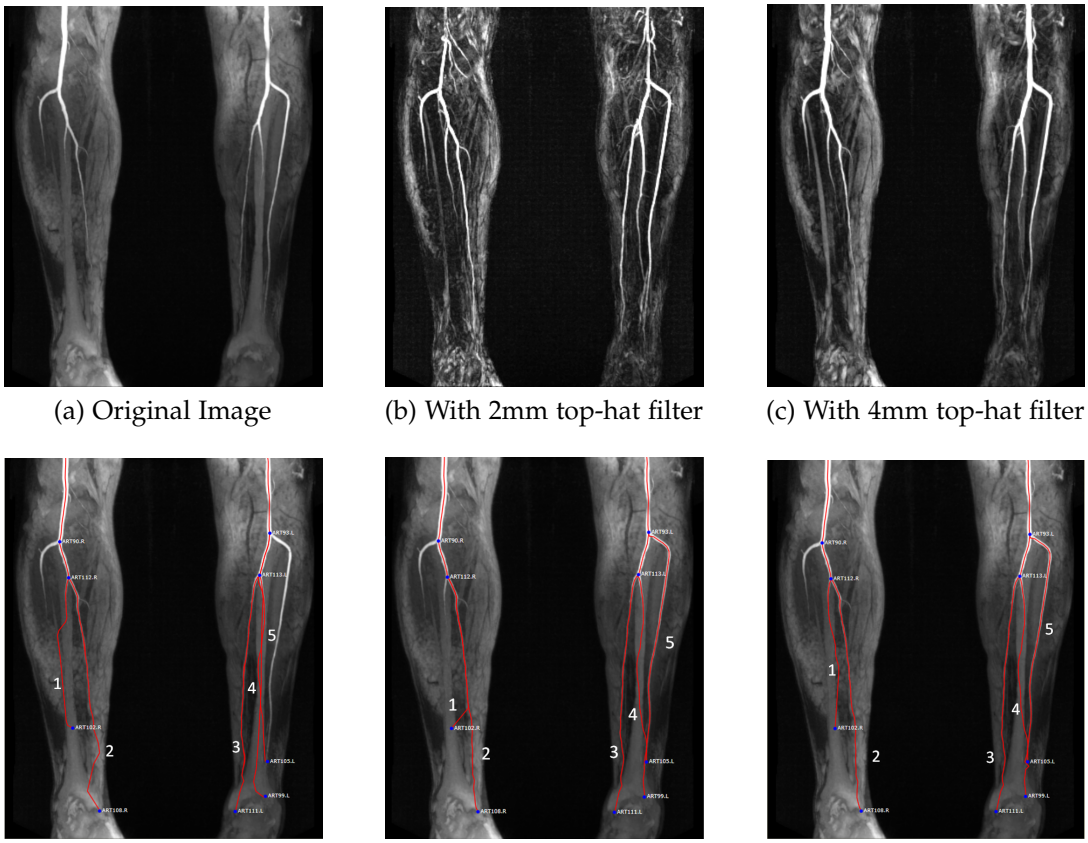


(b) Control, Sagittal View

(c) Using atlas, Coronal View

(d) Using atlas, Sagittal View

Figure 5.16: MIP images showing vessel tracking results for a scan of the head (Toshiba data), comparing a control algorithm (value of atlas assumed to be zero everywhere) with the method as described in this chapter. Note that filtering is not used for arteries in the head. Landmarks are marked with dots (blue), and tracked vessel paths are marked with lines (red). 1: R. internal carotid a., 2: L. internal carotid a., 3: R. vertebral a., 4: L. vertebral a., 5: R. external carotid a., 6: L. external carotid a., 7: R. common carotid a., 8: L. common carotid a., 9: R. subclavian a., 10: Brachiocephalic a., 11: L. subclavian a., 12: Aortic arch [TMVS Dataset ID: 2467]



(a) Original Image (b) With 2mm top-hat filter (c) With 4mm top-hat filter

(d) Control (e) With filter (f) With filter and atlas

Figure 5.17: Coronal view MIP images for an image of the lower legs (Siemens data). a), b) and c) show the original image, and then the same image with the two top-hat transform filters applied (4mm and 8mm). The filter removes the bright fat signal, and thus the vessels are not obscured on the MIP. The larger popliteal artery at the top of the image is plainly visible on the 8mm filtered image but has been suppressed on the 4mm filtered image. Images d), e) and f) show a comparison between the control, then with filtering added, and finally both the filtering and the vascular atlas. It should be noted that the atlas does not always improve tracking in the lower legs (see Table 5.3). The right anterior tibial artery was occluded in these images so was deemed not viable to track by the anatomist. Landmarks are marked with dots (blue), and tracked vessel paths are marked with lines (red). In f) all vessels are tracked correctly. 1: R. peroneal a., 2: R. posterior tibial a., 3: L. posterior tibial a., 4: L. peroneal a., 5: L. anterior tibial a. Images © Clinical Imaging University of Dundee [TMVS Dataset ID: 3458]

5.6 Vessel tracking from detected landmarks

The previous section shows tracking given manually placed landmarks. However, ultimately we would like the method to be fully automatic. In this section, we show results for the same datasets using automatically detected landmarks.

5.6.1 Method

The training data consists of the 14 stitched whole-body MRA datasets on which we validate the tracking, as well as a further 36 unstitched Siemens MRA datasets.

Datasets are pre-processed by smoothing and rescaling them to 2mm voxel^{-1} (from the initial resolution of 1mm voxel^{-1}).

A forest is trained using T = 80, D_T = 12 and F_T = 2500 (50% relative intensity features 50% gradient orientation features). A full set of parameters are given in Table 5.4. Feature offsets are selected using radial sampling. One pass of atlas location feature feedback using an *affine transformation* to map to the atlas is employed, such that although out-of-bag detection is employed, there is a small element of resubstitution as described previously. We assume this has negligible effect (see the results in section 2.4 for experimental justification).

After detection, landmarks with a probability $P_F(c|\boldsymbol{f})$ greater than $\tau_P = 0.2$ are considered to be positively detected, and tracking is initialised from these.

Parameter	Definition	Value
D_{Res}	Resolution at which detector is run	2mm voxel^{-1}
T	Number of trees in forest	80
D	Number of training datasets	50
D_T	Number of training datasets sampled per tree (bagging)	12
d_{max}	Maximum feature offset	50mm
F	Total number of possible features	1 $d_{sag}(v)$ + 65k intensity + 1800M grad. orient.
F_T	Number of features selected per tree	1 $d_{sag}(v)$ + 1250 intensity 1250 grad. orient.
$\sigma_{Sampling}$	Standard deviation of Gaussian weighting function for landmark samples	3.0mm
B_{Ratio}	Ratio of background to foreground training samples	5.0
B_{π_Ratio}	Ratio of background class to landmark class prior probability	400
w_{Node_min}	Minimum total weight of samples in a node for branch splitting, otherwise branch is terminated.	5.0
$w_{Node\_Split\_min}$	Minimum total weight of samples in smallest child node, otherwise branch is terminated.	2.0
d_{skip}	(Detection phase) Grid search interval	2 voxels
T_{min}	(Detection phase) Minimum number of trees to evaluate before forest shortcut may be deployed.	5 trees
$P_{Shortcut}$	(Detection phase) Minimum probability for forest shortcut.	0.15

Table 5.4: Empirically chosen parameter values for the random forest for whole-body MRA vascular landmark detection. This table can be directly compared with Table 2.1 for the original CT whole-body detector.

5.6.2 Results

Results are given in Table 5.5. Some pictorial results are shown in Figures 5.18, 5.19 and 5.20.

The tracking results from detected landmarks, whilst worse than tracking results from manually placed landmarks, still show much promise. A better detection algorithm could be achieved if more training data were available.

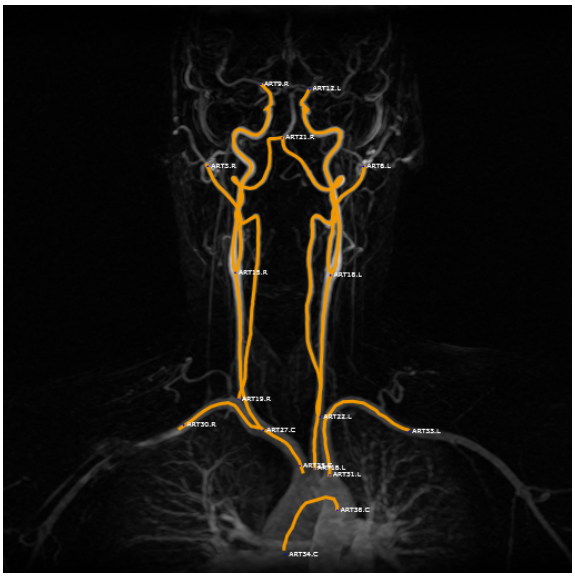
The run times, per dataset, are approximately 25 seconds for landmark detection and 10 minutes for vessel tracking. However, we have not started to optimise the tracking times. For instance, where there is vessel occlusion, the algorithm takes a long time to fail because we put very conservative limits on the maximum cost of the path and number of voxels visited. Also, the atlas creation has not been optimised. We re-load the ground truth each for each vessel in each dataset, and the code is not parallelised.

Table 5.5: Comparison of vessel tracking results from manually placed landmarks (semi-automatic tracking) and from detected landmarks (fully automatic tracking) in MRA datasets. n = number of test examples, GT = results from ground truth landmarks, Detected = results from detected landmarks. The results are expressed as the mean percentage of the vessel tracked successfully.

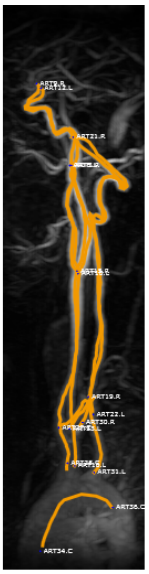
Artery	**n**	**GT**	**Detected**
		% *Tr*	**%** *Tr*
R External Carotid	13	71.2	73.8
L External Carotid	13	88.5	89.4
R Internal Carotid	13	83.2	82.7
L Internal Carotid	13	87.3	85.5
R Common Carotid	13	84.8	80.9
L Common Carotid	13	89.2	89.5
R Vertebral	12	74.0	73.9
L Vertebral	11	87.9	73.7
Brachiocephalic	14	100	100
R Subclavian	14	99.9	97.8
L Subclavian	14	100	100
Aortic Arch	14	100	100
Thoracic Aorta	13	100	100
Abdominal Aorta	12	100	100
Coeliac Trunk	13	99.5	68.5

Continued on next page

Artery	**n**	**GT**	**Detected**
		% *Tr*	% *Tr*
Superior Mesenteric	13	93.1	92.3
Right Main Renal	13	100	92.3
Left Main Renal	12	98.9	86.9
Inferior Mesenteric	10	43.1	21.7
R Common Iliac	12	91.0	90.7
L Common Iliac	12	91.3	91.4
R External Iliac	12	91.7	100
L External Iliac	12	100	91.7
R Common Femoral	12	100	89.4
L Common Femoral	12	100	100
R Profunda Femoris	12	98.6	77.7
L Profunda Femoris	12	97.3	72.5
R Superior Femoral	11	72.7	63.3
L Superior Femoral	12	75.0	81.8
R Popliteal	10	100	94.0
L Popliteal	12	96.1	95.1
R Anterior Tibial	8	73.2	71.1
L Anterior Tibial	9	92.4	70.3
R Tibioperoneal Trunk	11	98.3	82.5
L Tibioperoneal Trunk	11	100	100
R Peroneal	8	80.7	74.7
L Peroneal	9	85.7	62.9
R Posterior Tibial	11	75.2	74.9
L Posterior Tibial	9	72.6	35.2
Overall	460	90.3	84.3



(a) GT (Coronal)

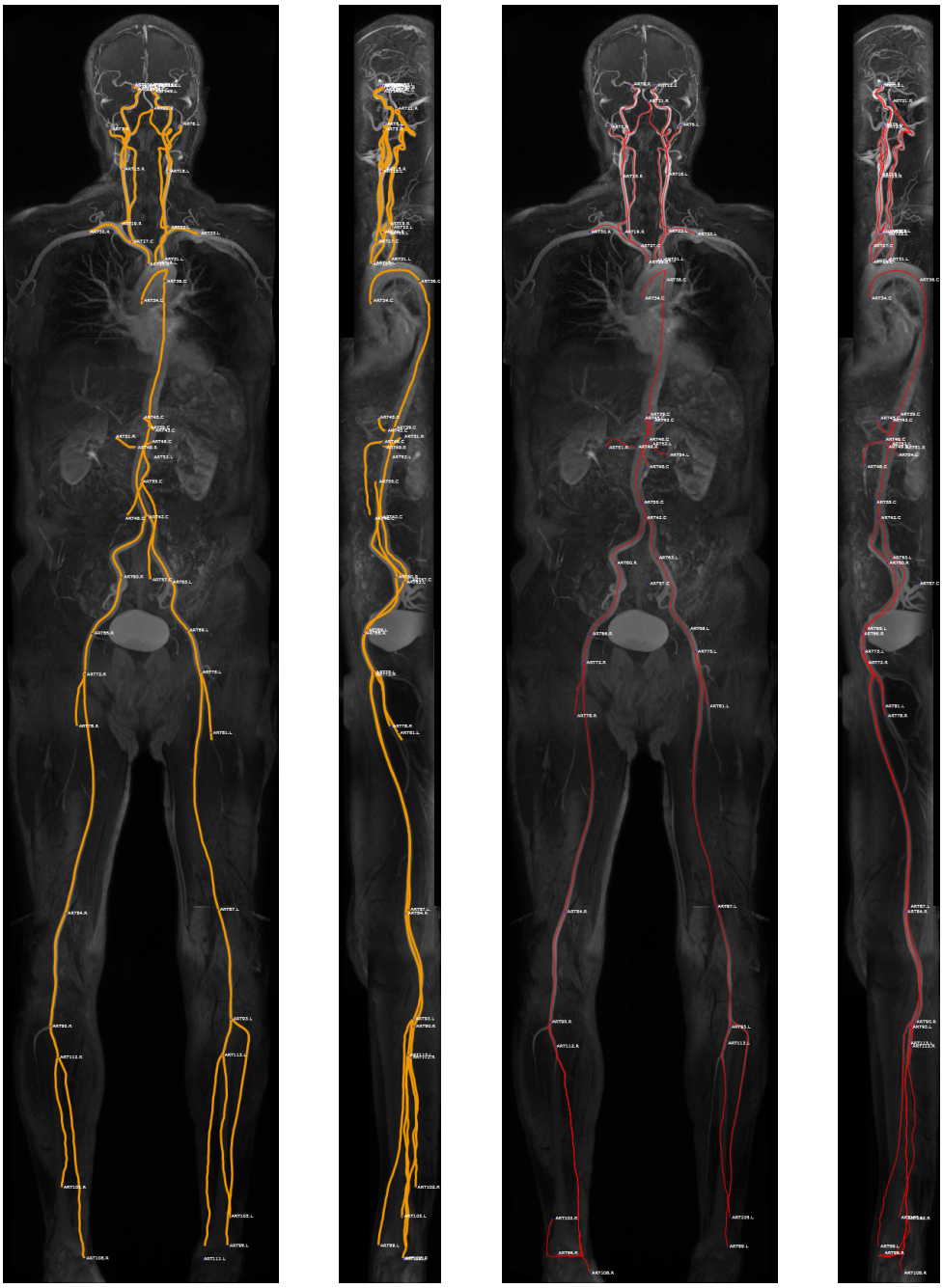


(b) GT (Sagittal)

(c) Detected (Coronal)

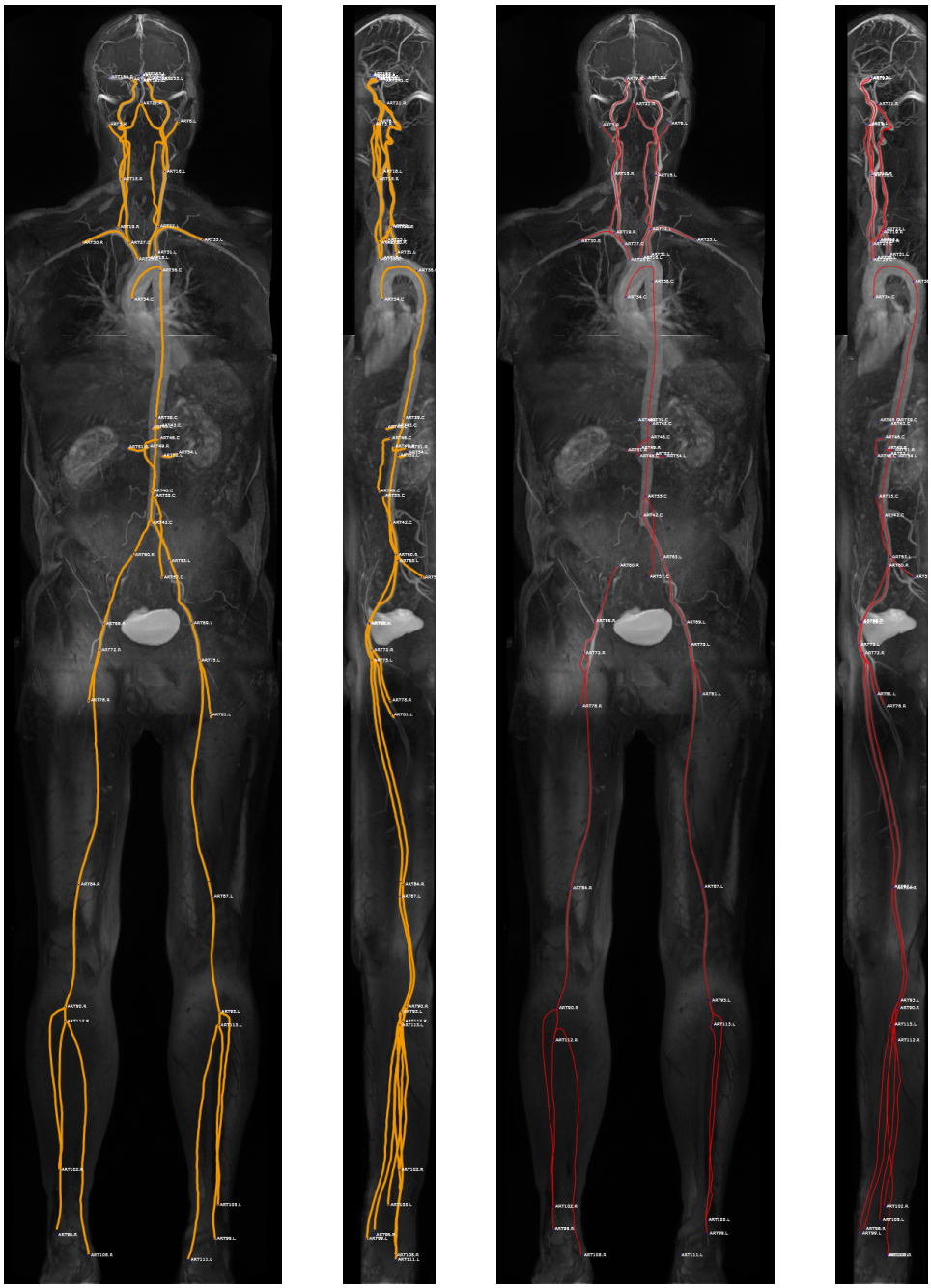
(d) Detected (Sagittal)

Figure 5.18: MIP images of the same Toshiba head dataset as shown in Figure 5.16, this time showing *fully* automatic vessel tracking results from *detected* landmarks. Overall the results are good, although there is some deviation from the vessel path on the left vertebral artery. The landmark ART21.R at the converging termini of the left and right vertebral arteries is also slightly mislocated. [TMVS Dataset ID: 2467]



(a) (b) (c) (d)

Figure 5.19: MIP images for a whole-body MRA dataset, showing fully automatic vessel tracking from detected landmarks. Figures 5.19a and 5.19b show the centre line ground truth in orange. Figures 5.19c and 5.19d show the detected centre lines in red. Most vessels are tracked satisfactorily. The inferior mesenteric artery has not been tracked. The right peroneal artery terminus landmark was marked as uncertain in the ground truth, so the poor detection and tracking is understandable. © Clinical Imaging University of Dundee [TMVS Dataset ID: 3458]



(a) (b) (c) (d)

Figure 5.20: MIP images for a whole-body MRA dataset, showing fully automatic vessel tracking from detected landmarks. Figures 5.20a and 5.20b show the centre line ground truth in orange. Figures 5.20c and 5.20d show the detected centre lines in red. Most vessels are tracked satisfactorily. However, the left posterior tibial artery has not been tracked, and the landmark ART72.R has been located on the wrong vessel, leading to the right common femoral, right deep femoral and right superior femoral arteries deviating slightly from their true paths. © Clinical Imaging University of Dundee [TMVS Dataset ID: 3600]

5.7 Discussion

5.7.1 Summary of our contribution

In summary, we have demonstrated a fully automatic method for identifying and tracking the major arteries that are present in an MRA dataset of the whole or part of the body. Landmarks at the bifurcations and termini of the major arteries of the body are located by a random forest detector. Subsequently a minimum path algorithm is used to track vessels one by one, incorporating knowledge of vessel diameter and path to guide the tracking process via two mechanisms. Firstly, *a priori* knowledge of the typical diameter of the vessel being tracked informs the scale of the morphological top-hat filter which is applied as a pre-processing step. Secondly, the two-point tracking algorithm employs a cost function based on intensity, shape and orientation information, the latter two of which are provided by a vascular atlas. This atlas (one per vessel) is registered by way of the detected landmark points, and is learnt from the (aggregated) ground truth centre lines in the training set.

The tracking method has been tested on a range of arteries of varying size and anatomical location, and it has been shown that there is a significant improvement when contextual information is used to inform tracking. For larger vessels, both the control method and the full method give excellent results. This is explained by the low tortuosity and bright contrast characteristic of large proximal vessels. For these vessels contextual knowledge is not required — but reassuringly there is no regression with its usage. For smaller, more peripheral vessels, there is an observable improvement when the algorithm is augmented with contextual information, and this is important because these vessels are often of clinical interest.

This is a nice demonstration of a possible application of automatically detected landmarks, which should be easily extendable to the venous system and to other modalities. There are many other existing algorithms which could be initialised or augmented with the aid of automatically detected landmark points.

5.7.2 Morphology versus curvature analysis: The effect of data resolution

In section 5.5.3, we described and compared the effects of two different filters on MRA data. The top-hat transform outperformed the Frangi vesselness filter,

because the latter was not sensitive to small vessels. Vesselness filtering often yielded zero response for e.g. the lower leg vessels, yet these were the vessels for which filtering was found to aid tracking the most (see Table 5.3).

This problem is referred to in the paper on scale-space by Florack *et al.* [188]. At scales close to the data resolution, the signal is effectively undersampled. In order for the scale-space approximation to be reasonable, the scale of the Gaussian kernel s must obey $s \gg s_0$, s_0 being the pixel scale. Namely, there is a requirement for the vessel cross-section to be significantly larger than the data scale. The whole-body MRA Siemens datasets used in this chapter have a scale of 1mm voxel^{-1} and for small vessels (see Figures 5.12 and 5.13) the Gaussian smoothing sigmas s may be as small as 1mm. Hence it is not a surprise that the estimation of the Hessian matrix for small vessels is poor.

By contrast, the top-hat transform is a non-linear operation which is sensitive to vessels of any cross-sectional size. It is possible to enhance even vessels which are one voxel thick, although at small scales both filter results will be compromised by spurious detail or noise.

5.8 Future work

5.8.1 Refinement of vessel tracking

Failure cases occur where vessels are significantly occluded or appear very dim. Ideas for improvements to the algorithm include:

- Making parameter values (for k_I, k_π and t_π) a function of the artery being tracked. These parameters could be defined as a function of vessel properties such as the (known) vessel tortuosity, and how predictable or variable its path is. In the case of tortuous vessels, we may want to discourage shortcuts across background voxels by increasing k_I (affects relative preference of foreground to background voxels). In the case of vessels with a highly unpredictable path, we may wish to decrease k_π and t_π so as to place less weight on the atlas prior.

- Replacing or supplementing the top-hat filter with either a band-pass top-hat filter or alternative filters such as the vessel-specific Frangi vesselness filter.

- Progressive refinement of the vascular atlas kernel density estimation by

making the kernel size a function of the known vessel diameter and the training data density at each point, and using anisotropic kernels [238].

The next step would be to acquire more data to truly evaluate the method. An acknowledged weakness of this work is that due to insufficient data, the parameter values were tuned on the same data for which the results are given. For more rigorous validation, results should be reported on a separate test set of data. In addition, the vascular landmark detector would be improved if more training data were available. As with any machine learning algorithm, significant quantities of data are required to train a good detector. It would also be interesting to introduce test data from a wider range of scanners and scan protocols.

5.8.2 Moving to multi-resolution detection

The observation was made in chapter 2 that landmarks on fine structures were not being detected with precision (see section 2.5.3.4). We proposed that the landmark features were simply at a finer scale than the forest operating resolution, and that the problem could be easily solved by running detection at a higher resolution. In this chapter, this is substantiated with a demonstration of landmark detection on narrow vessels at higher resolutions of 1mm voxel^{-1} (section 5.3) and 2mm voxel^{-1} (section 5.6).

However, running detection at a higher resolution entails a significantly longer detection time, of order $O(n^3)$ where n is the resolution, due to the cubic search space. Time could be saved by running the zeroth iteration at the lower 4mm voxel^{-1} resolution, graduating to a finer resolution for subsequent iterations.

Furthermore, the search space for iterations beyond the zeroth could be limited to the regions of interest which are discovered in the zeroth pass. For any given landmark, its region of interest would be defined *either* as those voxels with high probability in iteration zero, *or* as a neighbourhood to the landmark location. The neighbourhood could be defined either relative to the landmark in scan volume space, or in atlas space; its size and shape would be pre-learnt from the training data in the same way that mapping thresholds are learnt for atlas location autocontext.

Other authors have adopted a similar multi-resolution approach. For instance, Liu *et al.* [239] and Dikmen *et al.* [240] of Siemens use a coarse-to-fine detector cascade (8mm to 4mm to 2mm resolution), with progressive narrowing of the search space according to search spaces pre-defined relative to the landmark

location. In their approach, detectors are for single landmarks rather than using a single multi-landmark detector as we did, so the ordering of detector application also plays a role in the search strategy.

Additionally, we wonder if there is value in combining the results of detection at different resolutions. Figure 5.7 shows that the probability cloud is more spatially localised at lower resolution, but more sensitive to detail at higher resolution. Inspection of individual landmark errors shows that there is indeed ambiguity as to which resolution is best, see Figure 5.4. Classifier combination has already been explored for brain structure segmentation (see chapter 4); these ideas might have value for two random forest detectors trained at different resolutions and potentially also different feature sets.

5.8.3 Single-seed tracking of peripheral vessels

This chapter concentrated on the main, named arteries of the body. These arteries are present in most people, with much the same appearance. However, for the small vessels peripheral to these, which might have many configurations, a fully landmarked approach would be impractical.

For peripheral vessels, automatic tracking requires a single-seed tree tracking approach — the seed, or root, landmarks being those at the extremities of our current arterial tree. A number of such tree tracking algorithms were referenced in the prior art section. In the same way that we have used contextual information to aid two-point tracking, a tree-tracking algorithm could be aided by knowledge of the (typical) orientation and depth of the tree, the number of branching levels, the vessel diameter, and the rate of contrast attenuation.

Chapter 6

Conclusion

Abstract

To round off this thesis, we summarise the main conclusions from each of the preceding technical chapters, highlight the main themes, and reiterate the most promising ideas for future exploration.

6.1 Summary of research

6.1.1 Learned atlas location autocontext

How may spatial relationships between landmarks be exploited in a machine learning algorithm for anatomical landmark detection?

For this problem, we started with a random classification forest which predicted landmark locations on the basis of local neighbourhood HU intensities. In a classification paradigm, all voxels in a scan nominally have an equal chance of being selected as the location for any given landmark. There is no barrier to assigning anatomically implausible landmark configurations.

To incorporate spatial context into landmark detection, we developed a method of feeding back the results and hence iteratively refining them, which we term *atlas location autocontext*. This is analogous to the *autocontext* [28] mechanism used for segmentation problems. Using this mechanism, a sequence of n detectors is trained and applied in succession, with the predicted landmark locations from each detector fed as input to the next detector. The predicted landmark locations yield an estimated mapping to atlas space, and the estimated atlas x, y and z coordinate values for a given voxel v are used as machine learning features.

A number of rigid and deformable mappings were considered for registration to atlas space. Preliminary experiments showed that best registration accuracy was achieved by the deformable thin plate spline, with the underlying global transformation constrained to be a similarity transformation. The rigid mappings gave a significantly poorer fit. However, for the purpose of comparison, we tried atlas location features computed using both the thin plate spline (similarity-constrained) and an affine transformation.

The mappings gave a similar modest improvement in landmark localisation, as measured by mean error from the ground truth. The affine transformation gave slightly greater improvement in detection accuracy, as measured by AUC for a 30mm LROC. Improvement was observed after the first two rounds of feedback but thereafter the results were stable. It could also be seen qualitatively that the estimation of *uncertain* landmarks (see section 2.3.1.2) increased in credibility.

We attribute the unexpectedly good performance of the affine mapping to the fact that it has error-cancellation properties in the case of randomly distributed errors. Further, we note that the accuracy metric used in the preliminary experi-

ment was clouded by the inclusion of landmarks which are already well located by the original detector.

6.1.2 Gradient Orientation features

How may an existing machine learning algorithm for anatomical landmark detection in CT data be adapted for use in other imaging modalities?

For modalities other than CT, image intensities are uncalibrated and an anatomical landmark detector based on grey level intensities is much less powerful. Hence, in this work a set of features was sought which has greater invariance to transformations of the image intensities. We chose to investigate histograms of gradient orientations, since they are a well tried and tested image descriptor [79, 82, 87], which can be efficiently implemented using integral volumes.

Various aspects of gradient orientation histograms were explored: the dimensions of the cuboids over which the histogram is computed, the random sampling strategy for the cuboid offsets, the number of histogram bins, the plane in which the 2D gradient orientations are measured, noise thresholding and weighting schemes. In many aspects, the simplest scheme gave best performance. For instance, 8-binned histograms outperformed 16-binned histograms coupled with Gaussian weighting.

The optimised gradient orientation features were shown to better the performance of intensity features in MRI data, but not in CT data. In both modalities, a 50-50 mix of features gave slightly improved performance.

Finally, successful three-way cross-modality classification was performed with CT, MRI-T1 and MRI-T2 data. These detectors had worse performance than same modality classification but open up possibilities for situations dealing with a previously unseen data modality, or for pooling training data across modalities where data for the target modality is scarce.

6.1.3 Probabilistic fusion of MAS and RF classifiers

How may existing multi-atlas segmentation and machine learning classifiers be combined for the problem of brain region segmentation?

We considered the problem of brain parcellation from the MICCAI 2012 Grand

Challenge [1], which used a labelling protocol of 138 classes. This meant fine division of the brain into many regions, many of which appear similar. Hence, spatial context is important. Multi-atlas segmentation (MAS) classifiers are the established approach to brain parcellation, and we have an algorithm at TMVS [26] which is fast, robust and not far from state-of-the-art accuracy.

This algorithm originally used an expectation maximisation step for the purpose of post-processing the probabilistic results of the MAS classifier. However, we wondered if a random forest might provide a better alternative. We chose to approach the problem as one of *classifier combination* [138]. MAS and random forest classifiers were trained independently and the two sets of probabilities were simply combined by — in order to compare and contrast — both averaging and Bayesian multiplication.

Results showed that the random forest alone was a poor choice of classifier for this problem. However, when a forest classifier trained on very local image intensities ($\leq$15mm offset) was combined with the original MAS classifier, this gave significant improvement. Visual inspection showed that tissue types were better separated in regions of intricate cross-tissue boundaries. We presented results for two alternative forest classifiers which used longer range intensity features ($\leq$50mm offset) and gradient orientation features respectively. These classifiers performed better alone due to the richer spatial context, but gave rather less *independent* information compared to the MAS algorithm. The intuitive conclusion was that, where the ultimate goal is combination, the two base classifiers should be trained to be as different to one another as possible.

We discussed the application of class priors when working with random forests, pointing out that these could be applied at the *tree* level as was done for anatomical landmark detection, or more properly at the *forest* level as we have done with brain segmentation. We also discussed the importance of assessing the degree of independence when choosing a combination method (averaging or Bayesian multiplication). We note that in fact we applied a final empirical calibration step to all results which gave significant improvement, making the discussion of priors and of independence in combination methods somewhat redundant. It is proposed that in future both combination and calibration steps would be best replaced with a single machine learning step such as softmax regression.

6.1.4 Context-aware vessel tracking from detected landmarks

How may we develop a fully automated system for tracking and labelling the major arteries of the body in MRA datasets?

Many image analysis algorithms are *semi*-automated, requiring the user to place one or more manual seed points for initialisation. This chapter was a proof of concept for our claim that such manual seed points could be replaced with detected landmarks. We started with an existing two-point vessel tracking algorithm. Landmarks were defined for the whole arterial system such that tracking of major arteries could become fully automated.

In the first place, the anatomical landmark detector of earlier chapters was tuned for the purpose of angiography images and vascular landmarks. *Relative* intensities (i.e. relative to the intensity at the voxel of interest) were found to be outperform absolute intensities, for CT data as well as for MRI. On reflection this makes sense, since the intensity of contrast agent in scans is uncalibrated. We also required to move to a higher resolution than the previous detection resolution of 4mm voxel^{-1} in order to detect smaller vessels such as the coronary arteries. Finally, a radial pattern of sampling (i.e. uniform sampling with respect to the radial offset, yielding more features at short range) appeared to give better localisation of landmarks.

We then looked at how contextual information might be used to make the tracking more robust. This was done in two ways. Firstly, by morphological pre-processing of the image with a top-hat transform at a scale corresponding to the diameter of the vessel of interest. Secondly, by registering the ground truth centre lines to the detected landmarks using a thin plate spline mapping, and creating a vascular atlas — representing the spatial density of ground truth — which was used to influence the cost function for the minimum path tracking algorithm. Both of these mechanisms were made possible because we had knowledge of exactly where we were tracking from and to. Additionally, there were enough ancillary detected landmarks to allow registration of ground truth centre lines and thereby fit a vascular atlas, which would not be possible with just two vessel end points.

Promising results were achieved for fully automated tracking. This method could only improve given more data and more ground truth. Further, contextual information was shown to improve tracking even in the case of manually placed

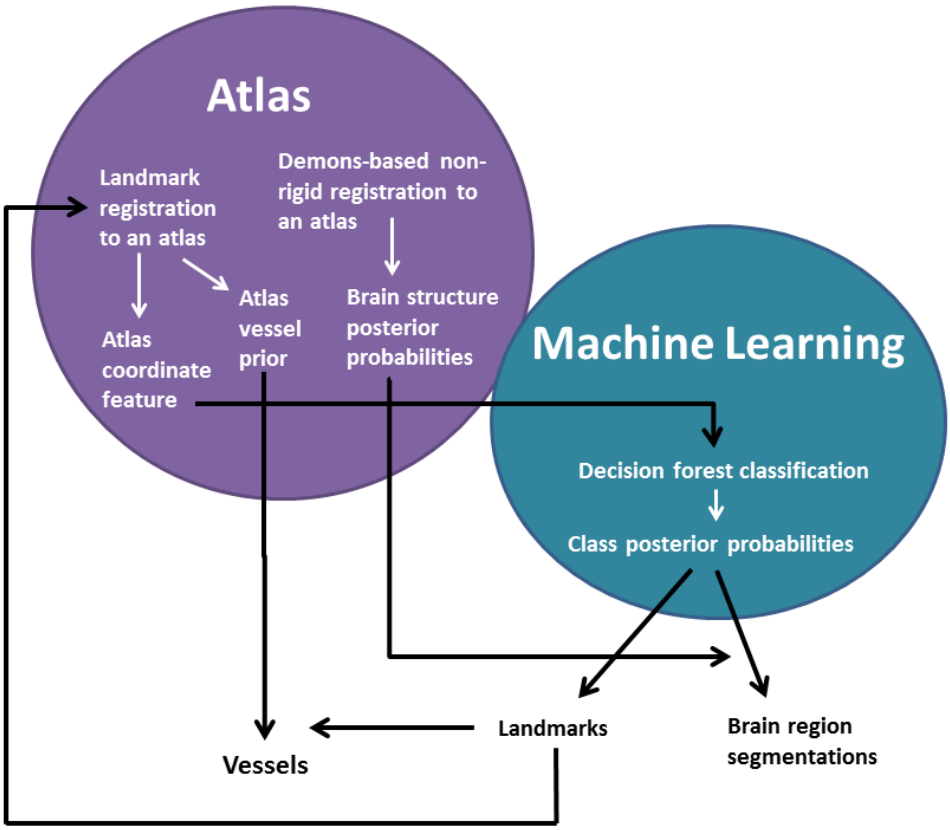


Figure 6.1: Diagram showing the relationships between the atlas-based and machine learning-based elements to each image analysis problem in this thesis.

landmarks, showing once again the importance of exploiting prior anatomical knowledge, be that atlas-based or otherwise.

6.2 Some higher-level observations

6.2.1 Combining atlas-based and feature-based information

We have explored the use of atlas-based and feature-based information for different anatomical structure detection problems. Figure 6.1 summarises the various relationships between the two. Ultimately both approaches gave useful and complementary information for all of the problems that we considered.

6.2.2 Mixing the old with the new

It was mentioned in the introduction (see section 1.4) that existing algorithms would be augmented or adapted where possible. In the end, we have showcased a

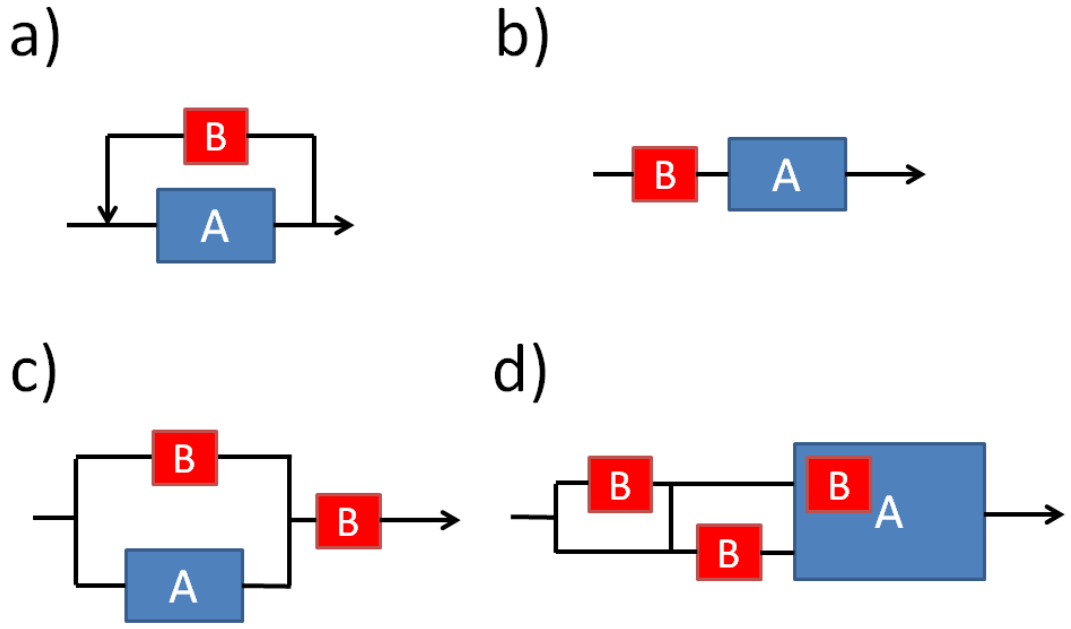


Figure 6.2: Diagram showing the architecture of the novel elements B and the pre-existing elements A. a) Atlas location autocontext: *Feedback* b) Gradient orientation features: *Input modification* c) Probabilistic fusion for segmentation: *Output modification* d) Vessel tracking from landmarks: *Novel inputs* (detected rather than manual landmarks), *Input modification* (data filtering) and *Algorithmic modification* (vascular atlas for tracking cost function).

range of different ways that a new element B may be introduced to a pre-existing algorithm A. Figure 6.2 shows the different architectures diagrammatically.

A nice side effect of this approach was that in all approaches, we were able to make direct comparisons of performance between the old algorithm A and the new algorithm $A + B$. This allowed easy assessment of the effectiveness of B.

On the other hand, our research directions were necessarily constrained by the path already trodden. One example is the use of a classification forest rather than a regression forest for landmark detection; the latter appears to be the more intuitive fit, and would fit seamlessly with the idea of atlas location autocontext.

6.2.3 The law of parsimony

The story of this thesis has been the triumph of simplicity over sophistication.

- Atlas autocontext worked better with a simple affine transform to atlas space than with a deformable thin plate spline transform.
- For gradient orientation features, the Gaussian weighting in orientation

space and noise level thresholding gave no significant benefit over using a small number of bins and uniform weighting.

- For vascular landmark detection, multi-size cuboidal features worked no better than single voxel intensities.
- For vessel enhancement, a simple morphological operation (the top-hat operation) outperformed the more complex Frangi vesselness filter.

According to Occam's razor (or the 'law of parsimony'), we chose the simplest effective method in each case. This shows the importance of proving the value of each level of complexity. We found that often, the simplest method was sufficient, or, there were other limiting factors which rendered the more complex method ineffective e.g. error in the detected landmarks for the spline, insufficient resolution of the data for the Frangi filter.

6.2.4 Features for modalities

The best choice of image features for the random forest is closely related to the image modality.

- For CT images, simple voxel intensities work well, although there is a slight benefit from adding gradient orientation features into the mix.
- For MRI and MRA images, gradient orientation features should be used, although there is a slight benefit from adding relative intensity features into the mix.
- CTA data should also be treated as uncalibrated data, with the use of relative intensities and — though the results have not been presented in this thesis — gradient orientation features.
- For unseen modalities, *unsigned* gradient orientation features should be used.

Each extra set of features carries a memory and computation time cost. The gradient orientation features require many times more memory than intensity features: one integral volume per histogram bin versus one integral volume in total. Hence, for any specific application there are trade-offs to be made.

6.3 Future research directions

Specific suggestions for future work are given at the end of each chapter. Here we reiterate the most promising ideas.

6.3.1 Acquiring more data

Machine learning algorithms thrive on data [241]. All of the methods in this thesis would benefit from more data. For instance, we use 278 training datasets in the whole-body CT anatomical landmark algorithm; this is by far the most data out of all experiments in this thesis. By contrast, 2000 datasets are used to train Siemens' ALPHA (Automatic Landmarking and Parsing of Human Anatomy) algorithm [239].

Currently we cannot claim to have found the limits of accuracy of any of our algorithms. Unfortunately, addition of more data also requires the painstaking and expensive manual creation of accompanying ground truth. However, the acquisition of data is definitely a priority for future work.

6.3.2 Improving and expanding atlas location features

In section 2.6.1 we outlined some interesting ideas for advancing the atlas mapping used for atlas location autocontext. Namely, by:

- Employing a *weighted* affine transformation, where the weights are informed by the known variances of the detection errors.
- Performing affine registration *by parts*, separating the registration of crudely separate articulating structures. Our initial suggestion is to split the body into: head, neck, thorax & abdomen, legs. The partial affine mappings would provide estimated atlas locations which could then be linked together with a thin plate spline to make a single mapping.

Further, we presented a number of interesting ideas which leverage the atlas space concept, in Table 2.6.2. These ideas range from expressing other image features (intensities, gradient orientations) in atlas space, to simply adjusting the operating resolution depending on the discovered scale of the person, in order to compensate for size differences between smaller and larger people.

6.3.3 Moving to multi-resolution landmark detection

In section 2.5.3.4 we discussed the causes of landmark detection imprecision, and concluded that some of our landmarks were situated on features which were not actually visible at the detection resolution. Correspondingly, in chapter 5 we moved to a higher resolution in order to detect vessels.

The recommendation going forward is to explore multi-resolution detection. The current setup, with multiple iterations, lends itself conveniently to a progressive stepped increase in resolution. We could simply run the first pass at 4mm voxel^{-1} and the second at 1mm voxel^{-1} or 2mm voxel^{-1}. Alternatively, we could start at an even lower resolution (perhaps 6mm voxel^{-1} or 8mm voxel^{-1}). The search space might be progressively narrowed as discussed, either by zoning in on those regions of high probability according to the previous iteration, or by using pre-defined search regions in atlas space.

Additionally, we propose that there may be value in mathematically combining the results of detection at different resolutions. Figure 5.7 shows that the probability cloud is more spatially localised at lower resolution, but more sensitive to detail at higher resolution. Inspection of individual landmark errors shows that there is indeed ambiguity as to which resolution is best, see Figure 5.4. The ideas of classifier combination which were explored for brain structure segmentation might have value for two random forest detectors trained at different resolutions and potentially also trained on different feature sets.

6.4 Final words

The contributions of this thesis have been highly practical, and there are some exciting results from both a research and commercial perspective. We have:

- Vastly reduced spurious results in anatomical landmark detection for CT data.
- Extended our landmark detection capability from CT to, in theory, any imaging modality.
- Demonstrated proof of concept for cross-modality landmark detection, in which the training and test modalities differ.
- Improved the accuracy of an existing brain gyrus segmentation algorithm.
- Developed a fully automated tracking and labelling system for the major arteries of the body in MRA scans.

In the process, we have challenged our ideas on what class prior probabilities mean in the context of a random forest, and even what the mechanism of a random forest really is. We have stressed the importance of validating every additional level of complexity in an algorithm. Our experimental hypothesis was proven wrong on more than one occasion, in which case we went to some length to explore why.

The merits of both atlas methods and machine learning, specifically random forests, have been reaffirmed and we have demonstrated many ways in which the two may co-exist in happy symbiosis.

Appendix A

List of anatomical landmarks

Below is a full list of landmark codes and descriptions. The original set of 127 landmarks is given first (as used in chapter 2 for the whole-body CT classifier).

The 'Head and neck' subset of 42 landmarks from the expanded (and redefined) list of 358 landmarks is then given (as used in chapter 3 for the MRI and CT head classifiers). This expanded list was defined at a later stage to supersede the original list. This list encompasses body regions not originally considered: hands, feet, skin surfaces and gender-specific organs, and also to remove or redefine those landmarks found to be poorly defined when analysing the results of the landmark detection classifier of chapter 2. In particular, landmarks were defined relative to the patient rather than relative to scanner space. Extra landmarks were also added for completeness (for instance, only the odd-numbered thoracic ribs and vertebrae were landmarked).

The landmark codes refer to the convenient abbreviations which were used when producing numerical results and labelling images (some of these are evident in the images shown in this thesis).

A.1 Original landmark list

Table A.1: Original list of landmarks. 'R' and 'L' refer to 'Right' and 'Left'. 'Post.' refers to 'Posterior'. 'CNS' = Central Nervous System. Note that the excretory system includes the kidney, the endocrine system includes the pancreas, the biliary system includes the liver, the 'Head' region comprises the head and the neck, and the 'Abdomen' region comprises the abdomen and the pelvis. There are no upper limb landmarks.

Description	Code	System	Region
Anterior arch of atlas (cervical vertebra I)	HNB1.A	Skeletal	Head
Superior tip of dens (cervical vertebra II)	HNB2.S	Skeletal	Head
Superior aspect of R eye globe	HNB3.S	Skeletal	Head
Superior aspect of L eye globe	HNB4.S	Skeletal	Head
Centre of R eye globe	HNB5.C	Skeletal	Head
Centre of L eye globe	HNB6.C	Skeletal	Head
Optic nerve attachment to right eye	HNB7.C	Brain & CNS	Head
Optic nerve attachment to left eye	HNB8.C	Brain & CNS	Head
Base of pituitary gland	HNB9.C	Brain & CNS	Head
Bifurcation of R common carotid artery	HNB10.S	Circulatory	Head
Bifurcation of L common carotid artery	HNB11.S	Circulatory	Head
Floor of R maxillary sinus	HNB12.I	Skeletal	Head
Floor of L maxillary sinus	HNB13.I	Skeletal	Head
Frontal horn of R lateral ventricle	HNB14.A	Brain & CNS	Head
Frontal horn of L lateral ventricle	HNB15.A	Brain & CNS	Head
Pineal gland	HNB16.C	Brain & CNS	Head
Centre of body of C3 vertebra	HNB17.C	Skeletal	Head
Centre of body of C4 vertebra	HNB18.C	Skeletal	Head

Continued on next page

Description	Code	System	Region
Centre of body of C5 vertebra	HNB19.C	Skeletal	Head
Centre of body of C6 vertebra	HNB20.C	Skeletal	Head
Centre of body of C7 vertebra	HNB21.C	Skeletal	Head
C1 spinous process, post. tip	HNB22.P	Skeletal	Head
C2 R bifid spinous process, post. tip	HNB23A.P	Skeletal	Head
C2 L bifid spinous process, post. tip	HNB23B.P	Skeletal	Head
C3 spinous process, post. tip	HNB24.P	Skeletal	Head
C4 spinous process, post. tip	HNB25.P	Skeletal	Head
C5 spinous process, post. tip	HNB26.P	Skeletal	Head
C6 spinous process, post. tip	HNB27.P	Skeletal	Head
C7 spinous process, post. tip	HNB28.P	Skeletal	Head
Bifurcation of trachea	THOR1.C	Respiratory	Thorax
Apex of R lung	THOR2.S	Respiratory	Thorax
Apex of L lung	THOR3.S	Respiratory	Thorax
Inferior angle of R scapula	THOR4.I	Skeletal	Thorax
Inferior angle of L scapula	THOR5.I	Skeletal	Thorax
Origin of L subclavian artery	THOR6.C	Circulatory	Thorax
Origin of L common carotid artery	THOR7.C	Circulatory	Thorax
Origin of brachiocephalic trunk	THOR8.C	Circulatory	Thorax
Bifurcation of brachiocephalic trunk	THOR9.C	Circulatory	Thorax
R coronary ostium	THOR10.C	Circulatory	Thorax
L coronary ostium	THOR11.C	Circulatory	Thorax
Aortic valve (centre of semilunar cusps)	THOR12.C	Circulatory	Thorax
Heart apex at epicardium (extremus in sagittal plane)	THOR13.L	Circulatory	Thorax
Heart apex at endocardium (extremus in sagittal plane)	THOR14.L	Circulatory	Thorax
Superior surface of sternal notch	THOR15.S	Skeletal	Thorax
Costophrenic angle of R lung	THOR16.R	Respiratory	Thorax
Costophrenic angle of L lung	THOR17.L	Respiratory	Thorax
R dome of diaphragm	THOR18.S	Respiratory	Thorax
L dome of diaphragm	THOR19.S	Respiratory	Thorax
Costal cartilage junction of 3rd R rib	THOR20.A	Skeletal	Thorax
Costal cartilage junction of 3rd L rib	THOR21.A	Skeletal	Thorax
Costal cartilage junction of 5th R rib	THOR22.A	Skeletal	Thorax
Costal cartilage junction of 5th L rib	THOR23.A	Skeletal	Thorax
Costal cartilage junction of 7th R rib	THOR24.A	Skeletal	Thorax
Costal cartilage junction of 7th L rib	THOR25.A	Skeletal	Thorax
Centre of body of T1 vertebra	THOR26.C	Skeletal	Thorax
Centre of body of T3 vertebra	THOR27.C	Skeletal	Thorax
Centre of body of T5 vertebra	THOR28.C	Skeletal	Thorax
Centre of body of T7 vertebra	THOR29.C	Skeletal	Thorax
Centre of body of T9 vertebra	THOR30.C	Skeletal	Thorax
Centre of body of T11 vertebra	THOR31.C	Skeletal	Thorax
Lateral extremus of R 3rd rib	THOR32.R	Skeletal	Thorax
Lateral extremus of L 3rd rib	THOR33.L	Skeletal	Thorax
Lateral extremus of R 5th rib	THOR34.R	Skeletal	Thorax
Lateral extremus of L 5th rib	THOR35.L	Skeletal	Thorax
Lateral extremus of R 7th rib	THOR36.R	Skeletal	Thorax
Lateral extremus of L 7th rib	THOR37.L	Skeletal	Thorax
Superior pole of R kidney	ABDO1.S	Excretory	Abdomen

Continued on next page

Description	Code	System	Region
Superior pole of L kidney	ABDO2.S	Excretory	Abdomen
Inferior pole of R kidney	ABDO3.I	Excretory	Abdomen
Inferior pole of L kidney	ABDO4.I	Excretory	Abdomen
Head of pancreas	ABDO5.M	Endocrine	Abdomen
Tip of tail of pancreas	ABDO6.L	Endocrine	Abdomen
Most inferior aspect of liver (R lobe)	ABDO7.I	Biliary	Abdomen
Posterior aspect of liver (R lobe)	ABDO8.P	Biliary	Abdomen
Origin of the coeliac trunk	ABDO9.C	Circulatory	Abdomen
Origin of the hepatic artery	ABDO10.C	Circulatory	Abdomen
Origin of the splenic artery	ABDO11.C	Circulatory	Abdomen
Origin of the superior mesenteric artery	ABDO12.C	Circulatory	Abdomen
Origin of R main renal artery	ABDO13.C	Circulatory	Abdomen
Origin of L main renal artery	ABDO14.C	Circulatory	Abdomen
Origin of R common iliac	ABDO15.C	Circulatory	Abdomen
Origin of L common iliac	ABDO16.C	Circulatory	Abdomen
Origin of R internal iliac artery	ABDO17.C	Circulatory	Abdomen
Origin of R external iliac artery	ABDO18.C	Circulatory	Abdomen
Origin of L internal iliac artery	ABDO19.C	Circulatory	Abdomen
Origin of L external iliac artery	ABDO20.C	Circulatory	Abdomen
Centre of body of L1 vertebra	ABDO21.C	Skeletal	Abdomen
Centre of body of L2 vertebra	ABDO22.C	Skeletal	Abdomen
Centre of body of L3 vertebra	ABDO23.C	Skeletal	Abdomen
Centre of body of L4 vertebra	ABDO24.C	Skeletal	Abdomen
Centre of body of L5 vertebra	ABDO25.C	Skeletal	Abdomen
Superior aspect of R iliac spine	ABDO26.S	Skeletal	Abdomen
Superior aspect of L iliac spine	ABDO27.S	Skeletal	Abdomen
R anterior superior iliac spine (ASIS)	ABDO28.A	Skeletal	Abdomen
L anterior superior iliac spine (ASIS)	ABDO29.A	Skeletal	Abdomen
Centre of symphasis pubis	ABDO30.C	Skeletal	Abdomen
Centre of head of R femur	ABDO31.C	Skeletal	Abdomen
Centre of head of L femur	ABDO32.C	Skeletal	Abdomen
R femur, greater trochanter (sup. aspect)	ABDO33.S	Skeletal	Abdomen
R femur, greater trochanter (lat. aspect)	ABDO34.L	Skeletal	Abdomen
L femur, greater trochanter (sup. aspect)	ABDO35.S	Skeletal	Abdomen
L femur, greater trochanter (lat. aspect)	ABDO36.L	Skeletal	Abdomen
Superior point of right sacro iliac joint	ABDO37.S	Skeletal	Abdomen
Inferior point of right sacro iliac joint	ABDO38.I	Skeletal	Abdomen
Superior point of left sacro iliac joint	ABDO39.S	Skeletal	Abdomen
Inferior point of left sacro iliac joint	ABDO40.I	Skeletal	Abdomen
Tip of the coccyx	ABDO41.I	Skeletal	Abdomen
Lateral epicondyle of R femur	LOWL1	Skeletal	Lower Limbs
Medial epicondyle of R femur	LOWL2	Skeletal	Lower Limbs
Lateral epicondyle of L femur	LOWL3	Skeletal	Lower Limbs
Medial epicondyle of L femur	LOWL4	Skeletal	Lower Limbs
Lateral condyle of R tibia	LOWL5	Skeletal	Lower Limbs
Medial condyle of R tibia	LOWL6	Skeletal	Lower Limbs
Lateral condyle of L tibia	LOWL7	Skeletal	Lower Limbs
Medial condyle of L tibia	LOWL8	Skeletal	Lower Limbs
Centre of head of R fibula	LOWL9	Skeletal	Lower Limbs
Centre of head of L fibula	LOWL10	Skeletal	Lower Limbs
Apex of R patella	LOWL11	Skeletal	Lower Limbs

Continued on next page

Description	Code	System	Region
Apex of L patella	LOWL12	Skeletal	Lower Limbs
Lateral malleolus of R fibula	LOWL13	Skeletal	Lower Limbs
Medial malleolus of R tibia	LOWL14	Skeletal	Lower Limbs
Lateral malleolus of L fibula	LOWL15	Skeletal	Lower Limbs
Medial malleolus of L tibia	LOWL16	Skeletal	Lower Limbs
Posterior aspect of R calcaneus	LOWL17	Skeletal	Lower Limbs
Posterior aspect of L calcaneus	LOWL18	Skeletal	Lower Limbs
Centre of dome of R talus	LOWL19	Skeletal	Lower Limbs
Centre of dome of L talus	LOWL20	Skeletal	Lower Limbs

A.2 Expanded landmark list: Head & neck subset

Table A.2: 'Head and neck' subset of the expanded landmark list. 'R' and 'L' refer to 'Right' and 'Left'. 'Post.' refers to 'Posterior'. 'CNS' = Central Nervous System. Note that the 'Surface' system refers to the skin surface.

Description	Code	System
Anterior arch of atlas	Atlas.A	Skeletal
Superior tip of dens / peg	Dens.S	Skeletal
Superior aspect of R eye globe	R.Eye.S	Skeletal
Superior aspect of L eye globe	L.Eye.S	Skeletal
Centre of R eye globe	R.Eye.C	Brain & CNS
Centre of L eye globe	L.Eye.C	Brain & CNS
Optic nerve attachment to R eye	R.CN2	Brain & CNS
Optic nerve attachment to L eye	L.CN2	Brain & CNS
Base of pituitary gland	PiGland.I	Brain & CNS
Tragus of R ear	R.Tragus	Surface
Tragus of L ear	L.Tragus	Surface
Opisthion (Posterior aspect of Foramen magnum)	Opisthion	Skeletal
Bifurcation of R common carotid artery	R.CoCarotid.Bi	Circulatory
Bifurcation of L common carotid artery	L.CoCarotid.Bi	Circulatory
Floor of R maxillary sinus	R.MaxSinus.I	Skeletal
Floor of L maxillary sinus	L.MaxSinus.I	Skeletal
Frontal horn of R lateral ventricle	R.LatVent.A	Brain & CNS
Frontal horn of L lateral ventricle	L.LatVent.A	Brain & CNS
Pineal gland	Pineal.C	Brain & CNS
Centre of body of C3	C3.C	Skeletal
Centre of body of C4	C4.C	Skeletal
Centre of body of C5	C5.C	Skeletal
Centre of body of C6	C6.C	Skeletal
Centre of body of C7	C7.C	Skeletal
Posterior arch (tubercle) of atlas	C1Tip.P	Skeletal
C2 R bifid spinous process, post. tip	R.TipC2.P	Skeletal
C2 L bifid spinous process, post. tip	L.TipC2.P	Skeletal
C3 spinous process, post. tip	C3Tip.P	Skeletal
C4 spinous process, post. tip	C4Tip.P	Skeletal
C5 spinous process, post. tip	C5Tip.P	Skeletal
C6 spinous process, post. tip	C6Tip.P	Skeletal

Continued on next page

Description	**Code**	**System**
C7 spinous process, post. tip	C7Tip.P	Skeletal
Body of hyoid bone (Anterior Aspect)	Hyoid.A	Skeletal
Glabella	Glabella	Surface
Nasion	Nasion	Surface
Acanthion	Acanthion	Skeletal
Mental protuberance of mandible	MandibleTub	Skeletal
Angle of R mandible	R.MandibleAng	Skeletal
Angle of L mandible	L.Mandible.Ang	Skeletal
Top of R ear attachment	R.TEA	Surface
Top of L ear attachment	L.TEA	Surface
Vertex of skull	Vertex	Skeletal

Appendix B

List of vascular landmarks and vessels

B.1 Vascular Landmarks

Table B.1 lists the 49 landmarks which were defined to describe the standard arterial tree in chapter 5, as shown in figure B.1.

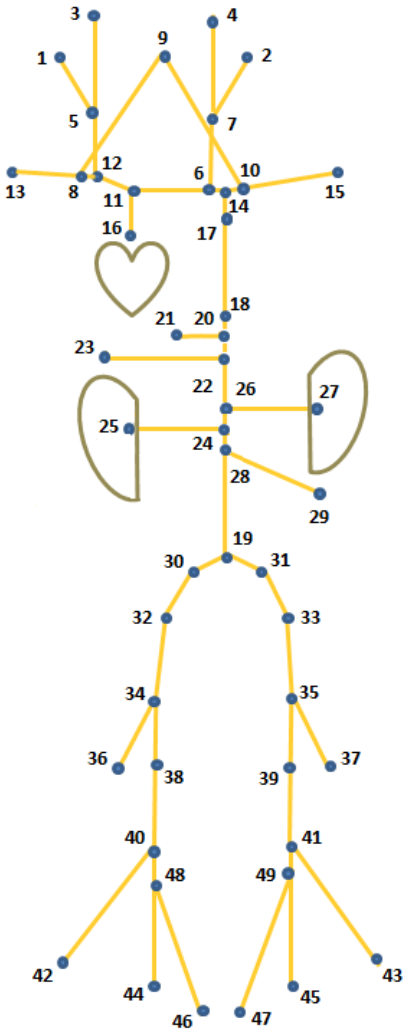


Figure B.1: Schematic of the standard arterial system. The blue dots represent landmarks, and the yellow lines represent arteries.

Table B.1: List of vascular landmarks.

Description	**Code**
1: Terminus of R external carotid a.	ART3.R
2: Terminus of L external carotid a.	ART6.L
3: Bifurcation of R internal carotid a.	ART9.R
4: Bifurcation of L internal carotid a.	ART12.L
5: Bifurcation of R common carotid a.	ART15.R
6: Origin of L common carotid a.	ART16.L
7: Bifurcation of L common carotid a.	ART18.L
8: Origin of right vertebral a.	ART19.R
9: Terminus of R and L vertebral arteries	ART21.R
10: Origin of L vertebral a.	ART22.L
11: Origin of brachiocephalic a.	ART25.C
12: Bifurcation of brachiocephalic a.	ART27.C
13: Terminus of R subclavian a.	ART30.R
14: Origin of L subclavian a.	ART31.L
15: Terminus of L subclavian a.	ART33.L
16: Origin of aortic arch	ART34.C
17: Terminus of aortic arch and origin of thoracic aorta	ART36.C
18: Terminus of thoracic aorta and origin of abdominal aorta	ART39.C
19: Bifurcation of abdominal aorta	ART42.C
20: Origin of coeliac a.	ART43.C
21: Terminus of coeliac a.	ART45.C
22: Origin of superior mesenteric a.	ART46.C
23: Terminus of superior mesenteric a.	ART48.C
24: Origin of R renal a.	ART49.R
25: Terminus of R renal a.	ART51.R
26: Origin of L renal a.	ART52.L
27: Terminus of L renal a.	ART54.L
28: Origin of inferior mesenteric a.	ART55.C
29: Terminus of inferior mesenteric a.	ART57.C
30: Bifurcation of R common Iliac a.	ART60.R
31: Bifurcation of L common Iliac a.	ART63.L
32: Terminus of R external iliac a. and origin of R common femoral a.	ART66.R
33: Terminus of L external iliac a. and origin of L common femoral a.	ART69.L
34: Branch point of R common femoral a.	ART72.R
35: Branch Point of L common femoral a.	ART75.L
36: Terminus of R profunda femoris a.	ART78.R
37: Terminus of L profunda femoris a.	ART81.L
38: Terminus of R superficial femoral a. and origin of R popliteal a.	ART84.R
39: Terminus of L superficial femoral a. and origin of L popliteal a.	ART87.L
40: Bifurcation of R politeal a.	ART90.R
41: Bifurcation of L popliteal a.	ART93.L
42: Terminus of R anterior tibial a.	ART96.R
43: Terminus of L anterior tibial a.	ART99.L
44: Terminus of R peroneal a.	ART102.R
45: Terminus of L peroneal a.	ART105.L
46: Terminus of R posterior tibial a.	ART108.R
47: Terminus of L posterior tibial a.	ART111.L
48: Bifurcation of R tibioperoneal trunk	ART112.R
49: Bifurcation of L tibioperoneal trunk	ART113.L

B.2 Vessels

Table B.2 lists the 39 vessels which were defined to describe the standard arterial tree in chapter 5, as shown in figure B.1.

Table B.2: List of vessels, including the landmarks located at the origin and terminus of each vessel.

Vessel	**Origin**	**Terminus**
R external carotid a.	ART15.R	ART3.R
L external carotid a.	ART18.L	ART6.L
R internal carotid a.	ART15.R	ART9.R
L internal carotid a.	ART18.L	ART12.L
R common carotid a.	ART27.C	ART15.R
L common carotid a.	ART16.L	ART18.L
R vertebral a.	ART19.R	ART21.R
L vertebral a.	ART22.L	ART21.R
Brachiocephalic	ART25.C	ART27.C
R. subclavian	ART27.C	ART30.R
L. subclavian	ART31.L	ART33.L
Aortic arch	ART34.C	ART36.C
Thoracic Aorta	ART36.C	ART39.C
Abdominal Aorta	ART39.C	ART42.C
Coeliac trunk	ART43.C	ART45.C
Superior mesenteric	ART46.C	ART48.C
R main renal	ART49.R	ART51.R
L main renal	ART52.L	ART54.L
Inferior mesenteric	ART55.C	ART57.C
R common iliac	ART42.C	ART60.R
L common iliac	ART42.C	ART63.L
R external iliac	ART60.R	ART66.R
L external iliac	ART63.L	ART69.L
R common femoral	ART66.R	ART72.R
L common femoral	ART69.L	ART75.L
R deep femoral	ART72.R	ART78.R
L deep femoral	ART75.L	ART81.L
R superficial femoral	ART72.R	ART84.R
L superficial femoral	ART75.L	ART87.L
R popliteal	ART84.R	ART90.R
L popliteal	ART87.L	ART93.L
R anterior tibial	ART90.R	ART96.R
L anterior tibial	ART93.L	ART99.L
R tibioperoneal trunk	ART90.R	ART112.R
L tibioperoneal trunk	ART93.L	ART113.L
R peroneal	ART112.R	ART102.R
L peroneal	ART113.L	ART105.L
R posterior tibial	ART112.R	ART108.R
L posterior tibial	ART113.L	ART111.L

Appendix C

List of brain structures

Table C.1 lists the brain structures which are segmented in chapter 4.

Table C.1: List of brain structures, split into cortical and non-cortical structures.

Description	**Category**
3rd ventricle	Non-cortical
4th Ventricle	Non-cortical
R Accumbens Area	Non-cortical
L Accumbens Area	Non-cortical
R Amygdala	Non-cortical
L Amygdala	Non-cortical
Brain Stem	Non-cortical
R Caudate	Non-cortical
L Caudate	Non-cortical
R Cerebellum Exterior	Non-cortical
L Cerebellum Exterior	Non-cortical
R Cerebellum White Matter	Non-cortical
L Cerebellum White Matter	Non-cortical
Right Cerebral White Matter	Non-cortical
Left Cerebral White Matter	Non-cortical
CSF	Non-cortical
Right Hippocampus	Non-cortical
L Hippocampus	Non-cortical
R Inf Lat Vent	Non-cortical
L Inf Lat Vent	Non-cortical
R Lateral Ventricle	Non-cortical
L Lateral Ventricle	Non-cortical
R Pallidum	Non-cortical
L Pallidum	Non-cortical
R Putamen	Non-cortical
L Putamen	Non-cortical
R Thalamus Proper	Non-cortical
L Thalamus Proper	Non-cortical
R Ventral DC	Non-cortical
L Ventral DC	Non-cortical
Optic Chiasm	Non-cortical
Cerebellar Vermal Lobules I-V	Non-cortical
Cerebellar Vermal Lobules VI-VII	Non-cortical
Cerebellar Vermal Lobules VIII-X	Non-cortical
R Basal Forebrain	Non-cortical
L Basal Forebrain	Non-cortical
R ACgG anterior cingulate gyrus	Cortical
L ACgG anterior cingulate gyrus	Cortical
R AIns anterior insula	Cortical
L AIns anterior insula	Cortical
R AnG angular gyrus	Cortical
L AnG angular gyrus	Cortical
R AOrG anterior orbital gyrus	Cortical
Left AOrG anterior orbital gyrus	Cortical
R Calc calcarine cortex	Cortical
Left Calc calcarine cortex	Cortical
R CO central operculum	Cortical

Continued on next page

Description	**Category**
L CO central operculum	Cortical
R Cun cuneus	Cortical
L Cun cuneus	Cortical
R Ent entorhinal area	Cortical
L Ent entorhinal area	Cortical
R FO frontal operculum	Cortical
L FO frontal operculum	Cortical
R FRP frontal pole	Cortical
L FRP frontal pole	Cortical
R FuG fusiform gyrus	Cortical
L FuG fusiform gyrus	Cortical
R GRe gyrus rectus	Cortical
L GRe gyrus rectus	Cortical
R IOG inferior occipital gyrus	Cortical
L IOG inferior occipital gyrus	Cortical
R ITG inferior temporal gyrus	Cortical
L ITG inferior temporal gyrus	Cortical
R LiG lingual gyrus	Cortical
L LiG lingual gyrus	Cortical
R LOrG lateral orbital gyrus	Cortical
L LOrG lateral orbital gyrus	Cortical
R MCgG middle cingulate gyrus	Cortical
L MCgG middle cingulate gyrus	Cortical
R MFC medial frontal cortex	Cortical
L MFC medial frontal cortex	Cortical
R MFG middle frontal gyrus	Cortical
L MFG middle frontal gyrus	Cortical
R MOG middle occipital gyrus	Cortical
L MOG middle occipital gyrus	Cortical
R MOrG medial orbital gyrus	Cortical
L MOrG medial orbital gyrus	Cortical
R MPoG postcentral gyrus medial segment	Cortical
L MPoG postcentral gyrus medial segment	Cortical
R MPrG precentral gyrus medial segment	Cortical
L MPrG precentral gyrus medial segment	Cortical
R MSFG superior frontal gyrus medial segment	Cortical
L MSFG superior frontal gyrus medial segment	Cortical
R MTG middle temporal gyrus	Cortical
L MTG middle temporal gyrus	Cortical
R OCP occipital pole	Cortical
L OCP occipital pole	Cortical
R OFuG occipital fusiform gyrus	Cortical
L OFuG occipital fusiform gyrus	Cortical
R OpIFG opercular part of the inferior frontal gyrus	Cortical
L OpIFG opercular part of the inferior frontal gyrus	Cortical
R OrIFG orbital part of the inferior frontal gyrus	Cortical
L OrIFG orbital part of the inferior frontal gyrus	Cortical
R PCgG posterior cingulate gyrus	Cortical
L PCgG posterior cingulate gyrus	Cortical
R PCu precuneus	Cortical
L PCu precuneus	Cortical

Continued on next page

Description	**Category**
R PHG parahippocampal gyrus	Cortical
L PHG parahippocampal gyrus	Cortical
R PIns posterior insula	Cortical
L PIns posterior insula	Cortical
R PO parietal operculum	Cortical
L PO parietal operculum	Cortical
R PoG postcentral gyrus	Cortical
L PoG postcentral gyrus	Cortical
R POrG posterior orbital gyrus	Cortical
L POrG posterior orbital gyrus	Cortical
R PP planum polare	Cortical
L PP planum polare	Cortical
R PrG precentral gyrus	Cortical
L PrG precentral gyrus	Cortical
R PT planum temporale	Cortical
L PT planum temporale	Cortical
R SCA subcallosal area	Cortical
L SCA subcallosal area	Cortical
R SFG superior frontal gyrus	Cortical
L SFG superior frontal gyrus	Cortical
R SMC supplementary motor cortex	Cortical
L SMC supplementary motor cortex	Cortical
R SMG supramarginal gyrus	Cortical
L SMG supramarginal gyrus	Cortical
R SOG superior occipital gyrus	Cortical
L SOG superior occipital gyrus	Cortical
R SPL superior parietal lobule	Cortical
L SPL superior parietal lobule	Cortical
R STG superior temporal gyrus	Cortical
L STG superior temporal gyrus	Cortical
R TMP temporal pole	Cortical
L TMP temporal pole	Cortical
R TrIFG triangular part of the inferior frontal gyrus	Cortical
L TrIFG triangular part of the inferior frontal gyrus	Cortical
R TTG transverse temporal gyrus	Cortical
L TTG transverse temporal gyrus	Cortical

Appendix D

Statement of technical contributions

The nature of this work is one of building on and improving existing TMVS solutions. Therefore, I state below clearly what I contributed.

This statement has been checked and approved by my industrial supervisor, Ian Poole.

D.1 Decision forest

The decision forest implementation that is used in this thesis is described in section 2.3.2, and in more detail in the paper by Dabbah *et al.* [25]. I was not involved in the initial implementation, however I became deeply involved in the optimisation and improvements across the codebase, particularly after two key scientists left TMVS.

Contributions are as follows:

D.1.1 Ground Truth

- I was involved in the analysis of intra-observer and inter-observer errors in the landmark ground truth data, and oversaw corrections.
- I incorporated and tested the ground truth for the expanded and improved landmark set (as used in chapter 3).

D.1.2 Atlas Location Autocontext

- I corrected the treatment of missing data values (to fit the strategy described in the paper).
- I was part of the discussion with Ian Poole and Sean Murphy through which the idea of using an atlas coordinate as a feedback feature was conceived.
- I removed the point atlas post-processing step (involving a regression equation), and developed a related method of fitting the atlas based on automatic finding of probability and distance thresholds (see section 2.3.3 for details).
- I debugged, unit-tested, and added multiple mapping types (spline types as well as two types of rigid mappings) to the initial implementation of the atlas feedback feature by Sean Murphy.
- I wrote the thin plate spline class (and associated unit tests and real data experiments).

D.1.3 Gradient Orientation Features

- I debugged, unit-tested, documented, and expanded the initial implementation of the gradient orientation feature by Mohammad Dabbah. The expansion consisted of moving from 2-dimensional to 3-dimensional feature boxes, adding measurement of 2D orientations in all three planes, fixing the data intensity normalisation, adding code to try Gaussian windowing and adding code to try noise level thresholding, and adding the ability to have both signed and unsigned features.
- I conceived the idea to train a classifier for unseen modalities.

D.1.4 Decision forest for segmentation

- I modified the classification forest framework for the task of segmentation (rather than landmark detection).
- I linked the (pre-existing) code for registration to the decision forest implementation and added the ability to combine the classifiers using simple addition and multiplication methods.
- I supervised Aneta Lisowska in the creation of a class to represent a compressed probability field (i.e. storing probability vectors for each voxel in a dataset, but retaining only the top n classes for each voxel), which we required for the segmentation task.
- During the brain segmentation work, I modified the class probability reweighting to be performed at the forest level rather than at the tree level. In retrospect, this is preferred as the technically correct method for landmark detection also.
- I wrote the code for empirical calibration of probabilities, following discussion with Ian Poole.
- I was involved in the analysis of intra-observer and inter-observer errors in the landmark ground truth data, and oversaw corrections.
- I incorporated and tested the ground truth for the expanded and improved landmark set (as used in chapter 3).

D.2 Vessel tracking algorithm

My contribution was to add context-awareness to a pre-existing two-point vessel tracking algorithm:

- I wrote a class that describes the arterial tree in terms of a graph of landmarks and their connecting vessels.
- I wrote a class to generate the sub-tree that is present in a dataset, given the landmarks which that dataset contains.
- I added code to create a vascular atlas for a given vessel.
- I modified the vessel tracking algorithm cost function to that described in chapter 5.
- I debugged the finite difference approximation to the Hessian matrix in the Frangi filter implementation.
- I wrote the vessel tracking pipeline, from detection of landmarks, through pre-processing top-hat filtering, through tracking using a vascular atlas, to output of relevant images, graphs and metrics.

References

[1] B. A. Landman and S. K. Warfield, eds., *Proc. MICCAI 2012 workshop on Multi-Atlas Labeling*. 2012.

[2] D. B. Evans, R. Elovainio, G. Humphreys, D. Chisholm, J. Kutzin, S. Russell, P. Saksena, and K. Xu, "Health systems financing: The path to universal coverage," tech. rep., World Health Organization, 2010.

[3] J. Beard, A. Officer, and A. Cassels, "World report on ageing and health," tech. rep., World Health Organization, 2015.

[4] D. M. Studdert, M. M. Mello, A. A. Gawande, T. K. Gandhi, A. Kachalia, C. Yoon, A. L. Puopolo, and T. A. Brennan, "Claims, errors, and compensation payments in medical malpractice litigation," *The New England Journal of Medicine*, vol. 354, no. 19, pp. 2024–33, 2006.

[5] G. N. Hounsfield, "Computed medical imaging. nobel lecture.," *Journal of Computer Assisted Tomography*, vol. 4, no. 5, pp. 665–74, 1980.

[6] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Medical Image Analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[7] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, J. H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren, R. P. Woods, J. J. Mann, and R. V. Parsey, "Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration," *NeuroImage*, vol. 46, no. 3, pp. 786–802, 2009.

[8] J. Ehrhardt, H. Handels, W. Plötz, and S. J. Pöppl, "Atlas-based recognition of anatomical structures and landmarks and the automatic computation of orthopedic parameters," *Methods of Information in Medicine*, vol. 43, no. 4, pp. 391–397, 2004.

[9] D. Liu and K. S. Zhou, "Anatomical landmark detection using nearest neighbor matching and submodular optimization," in *Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 7512 of *Lecture Notes in Computer Science*, pp. 393–401, 2012.

[10] M. Betke, H. Hong, D. Thomas, C. Prince, and J. P. Ko, "Landmark detection in the chest and registration of lung surfaces with an application to nodule registration," *Medical Image Analysis*, vol. 7, no. 3, pp. 265–281, 2003.

[11] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam, "Use of active shape models for locating structures in medical images," *Image and Vision Computing*, vol. 12, no. 6, pp. 355–365, 1994.

[12] T. F. Cootes, C. J. Taylor, D. Cooper, and J. Graham, "Active shape models - their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[13] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3D medical image segmentation: A review," *Medical Image Analysis*, vol. 13, no. 4, pp. 435–563, 2009.

[14] A. Criminisi, J. Shotton, and S. Bucciarelli, "Decision forests with long-range spatial context for organ localization in CT volumes," in *Proc. MICCAI workshop on Probabilistic Models in Medical Image Analysis (PMMIA)*, pp. 69–80, 2009.

[15] A. Montillo, J. S. J. Winn, J. Iglesias, D. Metaxas, and A. Criminisi, "Entangled decision forests and their application to semantic segmentation of CT images," in *Int. Conf. Information Processing in Medical Imaging (IPMI)*, vol. 6801 of *Lecture Notes in Computer Science*, pp. 184–196, 2011.

[16] R. Donner, B. H. Menze, H. Bischof, and G. Langs, "Global localization of 3D anatomical structures by pre-filtered Hough forests and discrete optimization," *Medical Image Analysis*, vol. 17, no. 8, pp. 1304–1314, 2013.

[17] Y. Guo, G. Wu, J. Jiang, and D. Shen, "Robust anatomical correspondence detection by hierarchical sparse graph matching," *IEEE Transactions on Medical Imaging*, vol. 32, no. 2, pp. 268–277, 2013.

[18] Y. Gao and D. Shen, "Context-aware anatomical landmark detection: Application to deformable model initialization in prostate CT images," in *Proc. MICCAI workshop on Machine Learning in Medical Imaging (MLMI)*, vol. 8679 of *Lecture Notes in Computer Science*, pp. 165–173, 2014.

[19] M. Schneider, S. Hirsch, B. Weber, G. Székely, and B. H. Menze, "Joint 3-D vessel segmentation and centerline extraction using oblique Hough forests with steerable filters," *Medical Image Analysis*, vol. 19, pp. 220–249, 2015.

[20] K. M. Cherry, B. Peplinski, L. Kim, S. Wang, L. Lu, W. Zhang, J. Liu, Z. Wei, and R. M. Summers, "Sequential Monte Carlo tracking of the marginal artery by multiple cue fusion and random forest regression," *Medical Image Analysis*, vol. 19, pp. 164–175, 2015.

[21] D. Zikic, B. Glocker, and A. Criminisi, "Atlas encoding by randomized forests for efficient label propagation," in *Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 8151 of *Lecture Notes in Computer Science*, pp. 66–73, 2013.

[22] H. Wang, S. R. Das, J. W. Suh, M. Altinay, J. Pluta, C. Craige, B. Avants, and P. A. Yushkevich, "A learning-based wrapper method to correct systematic errors in automatic image segmentation: Consistently improved performance in hippocampus, cortex and brain segmentation," *NeuroImage*, vol. 55, no. 3, pp. 968–985, 2011.

[23] R. Gauriau, R. Cuignet, D. Lesage, and I. Bloch, "Multi-organ localization combining global-to-local regression and confidence maps," in *Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 8675 of *Lecture Notes in Computer Science*, pp. 337–334, 2014.

[24] A. Sg2 Health Care Intelligence and Consulting, "RSNA simulcast," 2015.

[25] M. A. Dabbah, S. Murphy, H. Pello, R. Courbon, E. Beveridge, S. Wiseman, D. Wyeth, and I. Poole, "Detection and location of 127 anatomical landmarks in diverse CT datasets," in *Medical Imaging: Image Processing*, vol. 9034 of *Proc. SPIE*, p. 903415, 2014.

[26] S. Murphy, B. Mohr, Y. Fushimi, H. Yamagata, and I. Poole, "Fast, simple, accurate multi-atlas segmentation of the brain," in *Proc. Int. Workshop Biomedical Image Registration (WBIR)*, vol. 8545 of *Lecture Notes in Computer Science*, pp. 1–10, 2014.

[27] A. Kanitsar, R. Wegenkittl, and P. Felkel, "Postprocessing and visualization of peripheral CTA data in clinical environments," in *Proc. IEEE Visualization*, 2001.

[28] J. H. Morra, Z. Tu, L. G. Apostolova, A. E. Green, A. W. Toga, and P. M. Thompson, "Automatic subcortical segmentation using a novel contextual model," in *Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 5421 of *Lecture Notes in Computer Science*, pp. 194–201, 2008.

[29] K. Rohr, H. S. Stiehl, R. Sprengel, T. M. Buzug, J. Weese, and M. H. Kuhn, "Landmark-based elastic registration using approximating thin-plate splines," *IEEE Transactions on Medical Imaging*, vol. 20, no. 6, pp. 526–534, 2001.

[30] P. Hellier and C. Barillot, "Coupling dense and landmark-based approaches for nonrigid registration," *IEEE Transactions on Medical Imaging*, vol. 22, no. 2, pp. 217–227, 2003.

[31] M. Urschler, C. Zach, H. Ditt, and H. Bischof, "Automatic point landmark matching for regularizing nonlinear intensity registration: Application to thoracic CT images," in *Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 4191 of *Lecture Notes in Computer Science*, pp. 710–717, 2006.

[32] T. Polzin, J. Rühaak, R. Wernera, H. Handels, and J. Modersitzki, "Lung registration using automatically detected landmarks," *Methods of Information in Medicine*, vol. 53, no. 4, pp. 250–256, 2014.

[33] C. Dong, Y.-W. Chen, and C.-L. Lin, "Non-rigid registration with constraint of anatomical landmarks for assessment of locoregional therapy," in *IEEE International Conference on Information and Automation*, 2015.

[34] D. Han, Y. Gao, G. Wu, P.-T. Yap, and D. Shen, "Robust anatomical landmark detection for MR brain image registration," in *Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 8673 of *Lecture Notes in Computer Science*, pp. 186–193, 2014.

[35] D. Han, Y. Gao, G. Wua, P.-T. Yapa, and D. Shen, "Robust anatomical landmark detection with application to MR brain image registration," *Computerized Medical Imaging and Graphics*, vol. 46, pp. 277–290, 2015.

[36] N. Lay, N. Birkbeck, J. Zhang, and S. K. Zhou, "Rapid multi-organ segmentation using context integration and discriminative models," in *Int. Conf. Information Processing in Medical Imaging (IPMI)*, vol. 7917 of *Lecture Notes in Computer Science*, pp. 450–462, 2013.

[37] B. Ibragimov, B. Likar, F. PernuÅ¡, and T. Vrtovec, "A game-theoretic framework for landmark-based image segmentation," *IEEE Transactions on Medical Imaging*, vol. 31, no. 9, pp. 1761–1776, 2012.

[38] B. Ibragimov, B. Likar, F. PernuÅ¡, and T. Vrtovec, "Shape representation for efficient landmark-based segmentation in 3-D," *IEEE Transactions on Medical Imaging*, vol. 33, no. 4, pp. 861–874, 2014.

[39] C. Chen, W. Xie, J. Franke, P. A. Grutzner, L.-P. Nolte, and G. Zheng, "Automatic X-ray landmark detection and shape segmentation via data-driven joint estimation of image displacements," *Medical Image Analysis*, vol. 18, pp. 487–499, 2014.

[40] K. Rohr, "On 3D differential operators for detecting point landmarks," *Image & Vision Computing*, vol. 15, no. 3, pp. 219–233, 1997.

[41] S. Wörz and K. Rohr, "Localization of anatomical point landmark in 3D medical images by fitting 3D parametric intensity models," in *Int. Conf. Information Processing in Medical Imaging (IPMI)*, vol. 2732 of *Lecture Notes in Computer Science*, pp. 76–88, 2003.

[42] S. Wörz and K. Rohr, "Localization of anatomical point landmarks in 3D medical images by fitting 3D parametric intensity models," *Medical Image Analysis*, vol. 10, no. 1, pp. 41–58, 2006.

[43] J. Ashburner and K. J. Friston, "Unified segmentation," *NeuroImage*, vol. 26, no. 3, pp. 839–851, 2005.

[44] J. P. Thirion, "Image matching as a diffusion process: An analogy with Maxwell's demons," *Medical Image Analysis*, vol. 2, no. 3, pp. 243–260, 1998.

[45] J. E. Iglesias and N. Karssemeijer, "Robust initial detection of landmarks in film-screen mammograms using multiple FFDM atlases," *IEEE Transactions on Medical Imaging*, vol. 28, no. 11, pp. 1815–1824, 2009.

[46] P. Nair and A. Cavallaro, "3-d face detection, landmark localization, and registration using a point distribution model," *IEEE Transactions on Multimedia*, vol. 11, no. 4, pp. 611–623, 2009.

[47] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[48] C.-W. Wang, C.-T. Huang, M.-C. Hsieh, C.-H. Li, S.-W. Chang, W.-C. Li, R. Vandaele, R. Marée, S. Jodogne, P. Geurts, C. Chen, G. Zheng, C. Chu, H. Mirzaalian, G. Hamarneh, T. Vrtovec, and B. Ibragimov, "Evaluation and comparison of anatomical landmark detection methods for cephalometric X-ray images: A grand challenge," *IEEE Transactions on Medical Imaging*, vol. 34, no. 9, pp. 1890–1900, 2015.

[49] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[50] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 832–844, 1998.

[51] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine Learning*, pp. 1–22, 1999.

[52] P. Kontschieder, S. R. Bulo, H. Bischof, and M. Petillo, "Structured class-labels in random forests for semantic image labelling," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2011.

[53] I. Poole, "Optimal probabilistic relaxation labelling," in *Proc. British Machine Vision Conf. (BMVC)*, 1990.

[54] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3D brain image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1744–1757, 2010.

[55] Y. Gao and D. Shen, "Collaborative regression-based anatomical landmark detection," *Physics in Medicine and Biology*, vol. 60, pp. 9377–9401, 2015.

[56] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[57] J. R. Quinlan, *C4.5: Programs For Machine Learning*. Morgan Kauffman Publishers, 1992.

[58] A. G. Schwing, C. Zach, Y. Zheng, and M. Pollefeys, "Adaptive random forest - how many "experts" to ask before making a decision?," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1377–1384, 2011.

[59] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, pp. 567–585, 1989.

[60] D. T. Sandwell, "Biharmonic spline interpolation of GEOS-3 and SEASAT altimeter data," *Geophysical Research Letters*, vol. 14, no. 2, pp. 139–142, 1987.

[61] M. Lootus, T. Kadir, and A. Zisserman, "Vertebrae detection and labelling in lumbar MR images," in *Proc. 1st MICCAI Workshop on Computational Methods and Clinical Applications for Spine Imaging*, vol. 17 of *Lecture Notes in Computational Vision and Biomechanics*, pp. 219–230, 2014.

[62] J. Yao, T. Klinder, and S. Li, eds., *Proc. 1st MICCAI Workshop on Computational Methods and Clinical Applications for Spine Imaging*, vol. 17 of *Lecture Notes in Computational Vision and Biomechanics*, 2013.

[63] J. Yao, B. Glocker, T. Klinder, and S. Li, eds., *Proc. 2nd MICCAI Workshop on Computational Methods and Clinical Applications for Spine Imaging*, vol. 20 of *Lecture Notes in Computational Vision and Biomechanics*, 2014.

[64] T. Vrtovec, J. Yao, B. Glocker, T. Klinder, A. Frangi, G. Zheng, and S. LI, eds., *Proc. 3rd MICCAI Workshop on Computational Methods and Clinical Applications for Spine Imaging*, 2015.

[65] J. Piper, Y. Ikeda, Y. Fujisawa, Y. Ohnu, T. Yoshikawa, A. O'Neil, and I. Poole, "Objective evaluation of the correction by non-rigid registration of abdominal organ motion in low-dose 4D dynamic contrast-enhanced CT," *Physics in Medicine and Biology*, vol. 57, no. 6, pp. 1701–1715, 2012.

[66] A. C. Aitken, "On least squares and linear combinations of observations," in *Proc. of the Royal Society of Edinburgh*, vol. 55, pp. 42–48, 1935.

[67] B. Fischl, D. H. Salat, A. J. van der Kouwe, N. Makris, F. Segonne, B. T. Quinn, , and A. M. Dale, "Sequence-independent segmentation of magnetic resonance images," *NeuroImage*, vol. 23, pp. 69–84, 2004.

[68] A. Jog, A. Carass, S. Roy, D. L. Pham, and J. L. Prince, "MR image synthesis by contrast learning on neighborhood ensembles," *Medical Image Analysis*, vol. 24, pp. 63–76, 2015.

[69] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on feature distribution," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[70] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Medical Imaging*, vol. 24, no. 7, pp. 971–987, 2002.

[71] Y.-Y. Liu, M. Chen, H. Ishikawa, G. Wollstein, J. S. Schuman, and J. M. Rehg, "Automated macular pathology diagnosis in retinal OCT images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding," *Medical Image Analysis*, vol. 15, pp. 748–759, 2011.

[72] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[73] L. SÃ¸rensen, S. B. Shaker, and M. de Bruijne, "Texture classification in lung CT using local binary patterns," in *Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, no. 5241 in Lecture Notes in Computer Science, pp. 934–941, 2008.

[74] A. Oliver, X. Lladó, J. Freixenet, and J. Martí, "False positive reduction in mammographic mass detection using local binary patterns," in *Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, no. 4791 in Lecture Notes in Computer Science, pp. 286–293, 2007.

[75] F. Smeraldi, "Ranklets: orientation selective non-parametric features applied to face detection," in *Int. Conf. on Pattern Recognition (ICPR)*, vol. 3, pp. 379–382, 2002.

[76] M. Masotti and R. Campanini, "Texture classification using invariant ranklet features," *Pattern Recognition Letters*, vol. 23, pp. 1980–1986, 2008.

[77] M.-C. Yang, W. K. Moon, Y.-C. F. Wang, M. S. Bae, C.-S. Huang, J.-H. Chen, and R.-F. Chang, "Robust texture analysis using multi-resolution gray-scale invariant features for breast sonographic tumor diagnosis," *Medical Image Analysis*, vol. 32, no. 12, 2013.

[78] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.

[79] W. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition (FGR)*, pp. 296–301, 1995.

[80] R. K. McConnell, "Method of and apparatus for pattern recognition," 1986.

[81] W. T. Freeman, K. Tanaka, J. Ohta, and K. Kyuma, "Computer vision for computer games," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition (FGR)*, pp. 100–105, 1996.

[82] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 886–893, 2005.

[83] Q. Zhu, S. Avidan, M.-C. Yeh, and K.-T. Cheng, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006.

[84] B. Alefs, G. Eschemann, H. Ramoser, and C. Beleznai, "Road sign detection from edge orientation histograms," in *IEEE Intelligent Vehicles Symposium*, 2007.

[85] M. Bertozzi, A. Broggi, M. D. Rose, M. Felisa, A. Rakotomamonjy, and F. Suard, "A pedestrian detector using histograms of oriented gradients and a support vector machine classifier," in *IEEE Intelligent Transportation Systems Conference*, 2007.

[86] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. British Machine Vision Conf. (BMVC)*, 2008.

[87] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 1999.

[88] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[89] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Medical Imaging*, vol. 27, no. 10, 2005.

[90] S. Allaire, J. J. Kim, S. L. Breen, D. A. Jaffray, and V. Pekar, "Full orientation invariance and improved feature selectivity of 3D SIFT with application to medical image analysis," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR) workshops*, 2008.

[91] E. Haber and J. Modersitzki, "Intensity gradient based registration and fusion of multi-modal images," in *Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 726–733, 2006.

[92] J. Rühaak, L. König, M. Hallmann, N. Papenberg, S. Heldmann, H. Schumacher, and B. Fischer, "A fully parallel algorithm for multimodal image registration using normalized gradient fields," in *Proc. IEEE Int. Symp. on Biomedical Imaging (ISBI)*, 2013.

[93] M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, S. M. Brady, and J. A. Schnabel, "MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration," *Medical Image Analysis*, vol. 16, pp. 1423–1435, 2012.

[94] M. P. Heinrich, M. Jenkinson, B. W. Papiez, S. M. Brady, and J. A. Schnabel, "Towards realtime multimodal fusion for image-guided interventions using self-similarities," in *Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 187–194, 2013.

[95] Z. Li, D. Mahapatra, J. A. Tielbeek, J. Stoker, L. van Vliet, and F. M. Vos, "Image registration based on autocorrelation of local structure," *IEEE Transactions on Medical Imaging*, vol. 35, no. 1, pp. 63–75, 2015.

[96] Y. Ou, A. Sotiras, N. Paragios, and C. Davatzikos, "DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting," *Medical Image Analysis*, vol. 15, no. 4, pp. 622–639, 2011.

[97] H. Gudbjartsson and S. Patz, "The Rician distribution of noisy MRI data," *Magnetic Resonance in Medicine*, vol. 34, no. 6, pp. 910–914, 1995.

[98] M. Toews, L. Zöllei, and W. M. Wells, "Feature-based alignment of volumetric multi-modal images," in *Int. Conf. Information Processing in Medical Imaging (IPMI)*, pp. 25–36, Springer, Heidelberg, 2013.

[99] NeuroMorphoMetrics, inc., "NeuroMorphoMetrics segmentation protocol," Mar. 2015.

[100] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. M. Dale, "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, 2002.

[101] D. L. Collins, C. J. Holmes, T. M. Peters, and A. C. Evans, "Automatic 3D model-based neuroanatomical segmentation," *Human Brain Mapping*, vol. 3, no. 3, pp. 190–208, 1995.

[102] S. Sandor and R. Leahy, "Surface-based labeling of cortical anatomy using a deformable atlas," *IEEE Transactions on Medical Imaging*, vol. 16, no. 1, pp. 41–54, 1997.

[103] J. L. Lancaster, L. H. Rainey, J. L. Summerlin, C. S. Freitas, P. T. Fox, A. C. Evans, A. W. Toga, and J. C. Mazziotta, "Automated labelling of the human brain: A preliminary report on the development and evaluation of a forward-transform method," *Human Brain Mapping*, vol. 5, no. 4, pp. 238–242, 1997.

[104] B. M. Dawant, S. L. Hartmann, J.-P. Thirion, F. Maes, D. Vandermeulen, and P. Demaerel, "Automatic 3-D segmentation of internal structures of the head in MR images using a combination of similarity and free-form transformations. I. Methodology and validation on normal subjects," *IEEE Transactions on Medical Imaging*, vol. 18, no. 10, pp. 909–916, 1999.

[105] T. Gass, G. Székely, and O. Goksel, "Semi-supervised segmentation using multiple segmentation hypotheses from a single atlas (workshop on recognition techniques and applications in medical imaging)," in *Medical Computer Vision. Recognition and Applications in Medical Imaging*, vol. 7766 of *Lecture Notes in Computer Science*, pp. 29–37, 2012.

[106] J. C. Mazziotta, A. W. Toga, A. Evans, P. Fox, and J. Lancaster, "A probabilistic atlas of the human brain: Theory and rationale for its development," *NeuroImage*, vol. 2, pp. 89–101, 1995.

[107] J. Ashburner and K. J. Friston, "Voxel-based morphometry - the methods," *NeuroImage*, vol. 11, pp. 805–821, 2000.

[108] K. V. Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based tissue classification of MR images of the brain," *IEEE Transactions on Medical Imaging*, vol. 18, no. 10, pp. 897–908, 1999.

[109] T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer, "Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains," *NeuroImage*, vol. 21, no. 4, pp. 1428–1442, 2004.

[110] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *NeuroImage*, vol. 33, no. 1, pp. 115–126, 2006.

[111] P. Aljabar, R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy," *NeuroImage*, vol. 46, no. 3, pp. 726–738, 2009.

[112] J. M. Lötjönen, R. Wolz, J. R. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen, and D. Rueckert, "Fast and robust multi-atlas segmentation of brain magnetic resonance images," *NeuroImage*, vol. 49, pp. 2352–2365, 2010.

[113] M. Wu, C. Rosano, P. Lopez-Garcia, C. S. Carter, and H. J. Aizenstein, "Optimum template selection for atlas-based segmentation," *NeuroImage*, vol. 34, no. 4, pp. 1612–1618, 2007.

[114] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Medical Image Analysis*, vol. 24, pp. 205–219, 2015.

[115] X. Artaechevarria, A. M. noz Barrutia, and C. Ortíz de Solórzano, "Combination strategies in multi-atlas image segmentation: Application to brain MR data," *IEEE Transactions on Medical Imaging*, vol. 28, no. 8, pp. 1266–1277, 2009.

[116] M. R. Sabuncu, B. T. T. Yeo, K. V. Leemput, B. Fischl, and P. Golland, "A generative model for image segmentation based on label fusion," *IEEE Transactions on Medical Imaging*, vol. 29, no. 10, pp. 1714–1729, 2010.

[117] A. Akhondi-Asl and S. K. Warfield, "Evaluation of some STAPLE based fusion algorithms," in Landman and Warfield [1].

[118] A. J. Asman and B. A. Landman, "Formulating spatially varying performance in the statistical fusion framework," *IEEE Transactions on Medical Imaging*, vol. 31, no. 6, pp. 1326–1336, 2012.

[119] O. Commowick, A. Akhondi-Asl, and S. K. Warfield, "Estimating a reference standard segmentation with spatially varying performance parameters: Local MAP STAPLE," *IEEE Transactions on Medical Imaging*, vol. 31, no. 8, pp. 1593–1606, 2012.

[120] M. J. Cardoso, K. Leung, M. Modat, S. Keihaninejad, D. Cash, J. Barnes, N. C. Fox, and S. Ourselin, "STEPS: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation," *Medical Image Analysis*, vol. 17, pp. 671–684, 2013.

[121] A. Akhondi-Asl, L. Hoyte, M. E. Lockhart, and S. K. Warfield, "A logarithmic opinion pool based STAPLE algorithm for the fusion of segmentations with associated reliability weights," *IEEE Transactions on Medical Imaging*, vol. 33, no. 10, pp. 1997–2009, 2014.

[122] C. Ledig, R. A. Heckemann, A. Hammers, J. C. Lopez, V. F. Newcombe, A. Makropoulos, J. Lötjönen, D. K. Menon, and D. Rueckert, "Robust whole-brain segmentation: Application to traumatic brain injury," *Medical Image Analysis*, vol. 21, pp. 40–58, 2015.

[123] S. K. Warfield, K. H. Zhou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2005.

[124] F. van der Lijn, T. den Heijer, M. M. B. Breteler, and W. J. Niessen, "Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts," *NeuroImage*, vol. 43, no. 4, pp. 708–720, 2008.

[125] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 611–623, 2013.

[126] G. Wu, Q. Wang, D. Zhang, F. Nie, H. Huang, and D. Shen, "A generative probability model of joint label fusion for multi-atlas based brain segmentation," *Medical Image Analysis*, vol. 18, pp. 881–890, 2014.

[127] E. Smistad, T. L. Falch, M. Bozorgi, A. C. Elster, and F. Lindseth, "Medical image segmentation on GPUs â€“ a comprehensive review," *Medical Image Analysis*, vol. 20, pp. 1–18, 2015.

[128] A. J. Asman, Y. Huo, A. J. Plassard, and B. A. Landman, "Multi-atlas learner fusion: An efficient segmentation approach for large-scale data," *Medical Image Analysis*, vol. 26, pp. 82–91, 2015.

[129] P. Coupé, J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins, "Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation," *NeuroImage*, vol. 54, pp. 940–954, 2011.

[130] F. Rousseau, P. A. Habas, and C. Studholme, "A supervised patch-based approach for human brain labeling," *IEEE Transactions on Medical Imaging*, vol. 30, no. 10, pp. 1852–1862, 2011.

[131] W. Bai, W. Shi, C. Ledig, and D. Rueckert, "Multi-atlas segmentation with augmented features for cardiac MR images," *Medical Image Analysis*, vol. 19, pp. 98–109, 2015.

[132] Z. Wang, C. Donoghue, and D. Rueckert, "Patch-based segmentation without registration: application to knee MRI (workshop on machine learning in medical imaging)," in *Machine Learning in Medical Imaging*, no. 8184 in Lecture Notes in Computer Science, pp. 98–105, 2013.

[133] H. Wang and P. A. Yushkevich, "Multi-atlas segmentation without registration: A supervoxel-based approach," in *Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 8151 of *Lecture Notes in Computer Science*, pp. 535–542, 2013.

[134] X. Ren and J. Malik, "Learning a classification model for segmentation," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2003.

[135] J. E. Iglesias, C.-Y. Liu, P. M. Thompson, and Z. Tu, "Robust brain extraction across datasets and comparison with publicly available methods," *IEEE Transactions on Medical Imaging*, vol. 30, no. 9, pp. 1617–1634, 2011.

[136] Y. Meng, G. Li, Y. Gao, and D. Shen, "Automatic parcellation of cortical surfaces using random forests," in *Proc. IEEE Int. Symp. on Biomedical Imaging (ISBI)*, 2015.

[137] Z. Tu, S. Zheng, A. L. Yuille, A. L. Reiss, R. A. Dutton, A. D. Lee, A. M. Galaburda, I. Dinov, P. M. Thompson, and A. W. Toga, "Automated extraction of the cortical sulci based on a supervised learning approach," *IEEE Transactions on Medical Imaging*, vol. 26, no. 4, pp. 541–552, 2007.

[138] J. Kittler, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 2002.

[139] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*. Wiley-Interscience, 2004.

[140] S. Tulyakov, S. Jaeger, V. Govindaraju, and D. Doermann, "Review of classifier combination methods," in *Machine Learning in Document Analysis and Recognition*, vol. 90 of *Studies in Computational Intelligence*, pp. 361–386, Springer Berlin Heidelberg, 2008.

[141] J. A. Hoeting, "Bayesian model averaging: A tutorial," *Statistical Science*, pp. 382–401, 1999.

[142] A. P. Dawid, "The well-calibrated Bayesian," *Journal of the American Statistical Association*, vol. 77, no. 379, pp. 605–610, 1982.

[143] W. R. Crum, D. L. Hill, and D. J. Hawkes, "Information theoretic similarity measures in non-rigid registration," in *Int. Conf. Information Processing in Medical Imaging (IPMI)*, pp. 378–387, 2003.

[144] W. M. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," *Medical Image Analysis*, vol. 1, no. 1, pp. 35–51, 1995.

[145] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Mulmultimodal image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

[146] C. A. L. Bailer-Jones and K. Smith, "Combining probabilities," tech. rep., Max Planck Institute for Astronomy, 2011.

[147] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[148] T. SÃ¸rensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons," *Kongelige Danske Videnskabernes Selskab*, vol. 5, no. 4, pp. 1–34, 1948.

[149] H. Wang, B. Avants, and P. A. Yushkevich, "A combined joint label fusion and corrective learning approach," in Landman and Warfield [1].

[150] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *NeuroImage*, vol. 17, no. 2, pp. 825–41, 2002.

[151] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images. medical image analysis," *Medical Image Analysis*, vol. 5, no. 2, pp. 143–56, 2001.

[152] D. N. Greve and B. Fischl, "Accurate and robust brain image alignment using boundary-based registration," *NeuroImage*, vol. 48, no. 1, pp. 63–72, 2009.

[153] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios, "Dense image registration through MRFs and efficient linear programming," *Medical Image Analysis*, vol. 12, no. 6, pp. 731–741, 2008.

[154] M. J. Cardoso, M. Modat, R. Wolz, A. Melbourne, D. Cash, D. Rueckert, and S. Ourselin, "Geodesic information flows: Spatially-variant graphs and their application to segmentation and fusion," *IEEE Transactions on Medical Imaging*, vol. 34, no. 9, pp. 1976–1988, 2015.

[155] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, pp. 181–207, 2003.

[156] M. Gashler, C. Giraud-Carrier, and T. Martinez, "Decision tree ensemble: Small heterogeneous is better than large homogeneous," in *IEEE Int. Conf. on Machine Learning and Applications (ICMLA)*, pp. 900–905, 2008.

[157] K. Monteith, J. L. Carroll, K. Seppi, and T. Martinez, "Turning bayesian model averaging into bayesian model combination," *IEEE Int. Joint Conf. on Neural Networks (IJCNN)*, pp. 2657–2663, 2011.

[158] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[159] A. Vasilevskiy and K. Siddiqi, "Flux maximizing geometric flows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1565–1578, 2002.

[160] M. Schaap, T. van Walsum, L. Neefjes, C. Metz, E. Capuano, M. de Bruijne, and W. Niessen, "Robust shape regression for supervised vessel segmentation and its application to coronary segmentation in CTA," *IEEE Transactions on Medical Imaging*, vol. 30, no. 11, pp. 1974–1986, 2011.

[161] M. Piccinelli, A. Veneziani, D. A. Steinman, A. Remuzzi, and L. Antiga, "A framework for geometric analysis of vascular structures: Application to cerebral aneurysms," *IEEE Transactions on Medical Imaging*, vol. 28, no. 8, pp. 1141–1155, 2009.

[162] C. I. Christodoulou, C. S. Pattichis, M. Pantziaris, and A. Nicolaides, "Texture-based classification of atherosclerotic carotid plaques," *IEEE Transactions on Medical Imaging*, vol. 22, no. 7, pp. 902–912, 2003.

[163] M. Orkisz, M. Hernández-Hoyos, P. Douek, and I. Magnin, "Advances of blood vessel morphology analysis in 3D magnetic resonance images," *Machine Graphics and Vision*, vol. 9, no. 1/2, pp. 463–472, 2000.

[164] P. Felkel, R. Wegenkittl, and A. Kanitsar, "Vessel tracking in peripheral CTA datasets - an overview," in *EEE Spring Conf. Comput. Graphics*, pp. 232–239, 2001.

[165] J. S. Suri, K. Liu, L. Reden, and S. Laxminarayan, "A review on MR vascular image processing algorithms: Acquisition and prefiltering: Part I," *IEEE Transactions on Information Technology in Biomedicine*, vol. 6, no. 4, pp. 324–337, 2002.

[166] J. S. Suri, K. Liu, L. Reden, and S. Laxminarayan, "A review on MR vascular image processing: Skeleton versus nonskeleton approaches: Part II," *IEEE Transactions on Information Technology in Biomedicine*, vol. 6, no. 4, pp. 338–350, 2002.

[167] K. Bühler, P. Felkel, and A. L. Cruz, *Geometric methods for vessel visualization and quantificationâ€"A survey*. New York: Springer-Verlag, 2002.

[168] C. Kirbas and F. K. Quek, "Vessel extraction techniques and algorithms: a survey," in *Third IEEE Symposium on Bioinformatics and Bioengineering*, 2003.

[169] D. Lesage, E. D. Angelini, I. Bloch, and G. Funka-Lea, "A review of 3D vessel lumen segmentation techniques: Models, features and extraction schemes," *Medical Image Analysis*, vol. 13, pp. 819–845, 2009.

[170] C. A. Glasbey and G. W. Horgan, "Mathematical morphology," in *Image Analysis for the Biological Sciences*, ch. 5, Wiley, 1995.

[171] G. Matheron, *Random Sets and Integral Geometry*. Wiley, New York, 1975.

[172] J. Serra, *Image Analysis and Mathematical Morphology*. Academic Press, London, 1983.

[173] C. Ronse, "A lattice-theoretical morphological view on template extraction in images," *Journal of Visual Communication and Image Representation*, vol. 7, no. 3, pp. 273–295, 1996.

[174] P. Soille, *Morphological Image Analysis: Principles and Applications*. Springer, Berlin, Heidelberg, 2nd ed., 2003.

[175] C. Barat, C. Ducottet, and M. Jourlin, "Pattern matching using morphological probing," in *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 369–372, 2003.

[176] B. Naegel, N. Passat, and C. Ronse, "Grey-level hit-or-miss transformsâ€"part i: Unified theory," *Pattern Recognition*, vol. 40, pp. 635–647, 2007.

[177] B. N. Naegel, N. Passat, and C. Ronse, "Grey-level hit-or-miss transformsâ€"part ii: Application to angiographic image processing," *Pattern Recognition*, vol. 40, pp. 648–658, 2007.

[178] O. Merveille, H. Talbot, L. Najman, and N. Passat, "Tubular structure filtering by ranking orientation responses of path operators," in *European Conf. on Computer Vision (ECCV)*, vol. 8690 of *Lecture Notes in Computer Science*, pp. 203–218, 2014.

[179] Y. P. Du, D. L. Parker, and W. L. Davis, "Vessel enhancement filtering in three-dimensional MR angiography," *Journal of Magnetic Resonance Imaging*, vol. 5, no. 3, pp. 353–359, 1995.

[180] C. Lorenz, I.-C. Carlsen, T. Buzug, C. Fassnacht, and J. Weese, "Multi-scale line segmentation with automatic estimation of width, contrast and tangential direction in 2D and 3D medical images," in *CVRMed-MRCA'97*, vol. 1205 of *Lecture Notes in Computer Science*, pp. 233–242, 1997.

[181] Y. Sato, S. Nakajima, N. Shiraga, H. Atsumi, S. Yoshida, T. Koller, G. Gerig, and R. Kikinis, "Three-dimensional multi-scale line filter for segmentation and visualization of curvilinear structures in medical images," *Medical Image Analysis*, vol. 2, no. 2, pp. 143–168, 1998.

[182] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 1496 of *Lecture Notes in Computer Science*, pp. 130–137, 1998.

[183] K. Krissian, G. Malandain, N. Ayache, R. Vaillant, and Y. Trousset, "Model-based detection of tubular structures in 3D images," *Computer vision and image understanding*, vol. 80, no. 2, pp. 130–171, 2000.

[184] Q. Li, S. Sone, and K. Doi, "Selective enhancement filters for nodule, vessels, and airway walls in two- and three-dimensional CT scans," *Medical Physics*, vol. 30, no. 8, pp. 2040–2051, 2003.

[185] R. Manniesing, M. A. Viergever, and W. J. Niessen, "Vessel enhancing diffusion: A scale space representation of vessel structures," *Medical Image Analysis*, vol. 10, pp. 815–825, 2006.

[186] C. Xiao, M. Staring, Y. Wang, D. P. Shamonin, and B. C. Stoel, "Multiscale bi-Gaussian filter for adjacent curvilinear structures detection with application to vasculature images," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 174–188, 2013.

[187] J. Yang, S. Ma, Q. Sun, W. Tan, M. Xu, N. Chen, and D. Zhao, "Improved Hessian multiscale enhancement filter," *Bio-Medical Materials and Engineering*, vol. 24, no. 6, pp. 3267–3275, 2014.

[188] L. M. J. Florack, B. M. ter Haar Romeny, J. J. Koenderink, and M. A. Viergever, "Scale and the differential structure of images," *Image & Vision Computing*, vol. 10, no. 6, pp. 376–388, 1992.

[189] T. Lindeberg, "Scale-space for discrete signals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 3, pp. 234–254, 1990.

[190] T. Lindeberg, "Edge detection and ridge detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 117–156, 1998.

[191] B. E. Chapman and D. L. Parker, "3D multi-scale vessel enhancement filtering based on curvature measurements: application to time-of-flight MRA," *Medical Image Analysis*, vol. 9, pp. 191–208, 2005.

[192] C. Bauer and H. Bischof, "A novel approach for detection of tubular objects and its application to medical image analysis," *Pattern Recognition*, pp. 163–172, 2008.

[193] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow," *IEEE Transactions on Image Processing*, vol. 7, no. 3, pp. 359–369, 1998.

[194] M. W. K. Law and A. C. S. Chung, "Three dimensional curvilinear structure detection using optimally oriented flux," in *European Conf. on Computer Vision (ECCV)*, vol. 5305 of *Lecture Notes in Computer Science*, pp. 368–382, 2008.

[195] R. Moreno and O. Smedby, "Gradient-based enhancement of tubular structures in medical images," *Medical Image Analysis*, vol. 26, pp. 19–29, 2015.

[196] F. Zana and J.-C. Klein, "Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation," *IEEE Transactions on Image Processing*, vol. 10, no. 7, pp. 1010–1019, 2001.

[197] A. M. MendonÃ§a and A. Campilho, "Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction," *IEEE Transactions on Medical Imaging*, vol. 25, no. 9, pp. 1200–1213, 2006.

[198] A. Dufour, O. Tankyevych, B. Naegel, H. Talbot, C. Ronse, J. Baruthio, P. Dokládal, and N. Passat, "Filtering and segmentation of 3D angiographic data: Advances based on mathematical morphology," *Medical Image Analysis*, vol. 17, pp. 147–164, 2013.

[199] X. Qian, M. P. Brennan, D. P. Dione, W. L. Dobrucki, M. P. Jackowski, C. K. Breuer, A. J. Sinusas, and X. Papademetris, "A non-parametric vessel detection method for complex vascular structures," *Medical Image Analysis*, vol. 13, pp. 49–61, 2009.

[200] P. T. Truc, M. A. Khan, Y.-K. Lee, S. Lee, and T.-S. Kim, "Vessel enhancement filter using directional filter bank," *Computer Vision and Image Understanding*, vol. 113, pp. 101–112, 2009.

[201] W. T. Freeman and E. Adelson, "The design and use of steerable filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 9, 1991.

[202] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[203] O. Wink, W. J. Niessen, and M. A. Viergever, "Minimum cost path determination using a simple heuristic function," in *IEEE International Conference on Pattern Recognition*, vol. 3, pp. 998–1001, 2000.

[204] H. Li and A. Yezzi, "Vessels as 4-D curves: global minimal 4-D paths to extract 3-D tubular surfaces and centerlines," *Medical Image Analysis*, vol. 26, pp. 1213–1223, 2007.

[205] C. T. Metz, M. Schaap, T. van Walsum, and W. J. Niessen, "Two point minimum cost path approach for CTA coronary centerline extraction," *The Insight Journal*, p. 639, 2008.

[206] F. Benmansour and L. D. Cohen, "Tubular structure segmentation based on minimal path method and anisotropic enhancement," *International Journal of Computer Vision*, vol. 92, no. 2, pp. 192–210, 2011.

[207] S. D. Olabarriaga, M. Breeuwer, and W. J. Niessen, "Minimum cost path algorithm for coronary artery central axis tracking in CT images," in *Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 2879 of *Lecture Notes in Computer Science*, pp. 687–694, 2003.

[208] Y. Sun, "Automated identification of vessel contours in coronary arteriograms by an adaptive tracking algorithm," *IEEE Transactions on Medical Imaging*, vol. 8, no. 1, pp. 78–88, 1989.

[209] K. Haris, S. N. Efstratiadis, N. Maglaveras, C. Pappas, J. Gourassas, and G. Louridas, "Model-based morphological segmentation and labeling of coronary angiograms," *IEEE Transactions on Medical Imaging*, vol. 18, no. 10, 1999.

[210] R. Manniesing, B. Velthuis, M. van Leeuwen, I. van der Schaaf, P. van Laar, and W. Niessen, "Level set based cerebral vasculature segmentation and diameter quantification in CT angiography," *Medical Image Analysis*, vol. 10, pp. 200–214, 2006.

[211] J. Lee, P. Beighley, E. Ritman, and N. Smith, "Automatic segmentation of 3D micro-CT coronary vascular images," *Medical Image Analysis*, vol. 11, pp. 630–647, 2007.

[212] N. Flasque, M. Desvignes, J.-M. Constans, and M. Revenu, "Acquisition, segmentation and tracking of the cerebral vascular tree on 3D magnetic resonance angiography images," *Medical Image Analysis*, vol. 5, no. 3, pp. 173–183, 2001.

[213] Y. Fridman, S. M. Pizer, S. Aylward, and E. Bullitt, "Extracting branching tubular object geometry via cores," *Medical Image Analysis*, vol. 8, pp. 169–176, 2004.

[214] N. Passat, C. Ronse, J. Baruthio, J.-P. Armspach, C. Maillot, and C. Jahn, "Region-growing segmentation of brain vessels: An atlas-based automatic approach," *Journal of Magnetic Resonance Imaging*, vol. 21, no. 715-725, 2005.

[215] U. Jandt, D. Schäfer, M. Grass, and V. Rasche, "Automatic generation of 3D coronary artery centerlines using rotational X-ray angiography," *Medical Image Analysis*, vol. 13, pp. 846–858, 2009.

[216] C. Bauer, T. Pock, E. Sorantin, H. Bischof, and R. Beichel, "Segmentation of interwoven 3D tubular tree structures utilizing shape priors and graph cuts," *Medical Image Analysis*, vol. 14, pp. 172–184, 2010.

[217] B. Bouraoui, C. Ronse, J. Baruthio, N. Passat, and P. Germain, "3D segmentation of coronary arteries based on advanced mathematical morphology techniques," *Medical Image Analysis*, vol. 34, pp. 377–387, 2010.

[218] C. Zhou, H.-P. Chan, A. Chughtai, S. Patel, L. M. Hadjiiski, J. Wei, and E. A. Kazerooni, "Automated coronary artery tree extraction in coronary CT angiography using a multiscale enhancement and dynamic balloon tracking (MSCAR-DBT) method," *Medical Image Analysis*, vol. 36, pp. 1–10, 2012.

[219] S. Cetin, A. Demir, A. Yezzi, M. Degertekin, , and G. Unal, "Vessel tractography using an intensity based tensor model with branch detection," *IEEE Transactions on Medical Imaging*, vol. 32, no. 2, pp. 348–363, 2013.

[220] D. A. B. Oliveira, L. Leal-Taixé, R. Q. Feitosa, and B. Rosenhahn, "Automatic tracking of vessel-like structures from a single starting point," *Computerized Medical Imaging and Graphics*, vol. 47, pp. 1–15, 2016.

[221] M. Schaap, R. Manniesing, I. Smal, T. van Walsum, A. van der Lugt, and W. Niessen, "Bayesian tracking of tubular structures and its application to carotid arteries in CTA," in *Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 4792 of *Lecture Notes in Computer Science*, pp. 562–570, 2007.

[222] O. Friman, M. Hindennach, C. Kühnel, and H.-O. Peitgen, "Multiple hypothesis template tracking of small 3D vessel structures," *Medical Image Analysis*, vol. 14, pp. 160–171, 2010.

[223] X. Wang, T. Heimann, P. Lo, M. Sumkauskaite, M. Puderbach, M. de Bruijne, H. P. Meinzer, and I. Wegner, "Statistical tracking of tree-like tubular structures with efficient branching detection in 3D medical image data," *Physics in Medicine and Biology*, vol. 57, pp. 5325–5342, 2012.

[224] P. J. Yim, P. L. Choyke, and R. M. Summers, "Gray-scale skeletonization of small vessels in magnetic resonance angiography," *IEEE Transactions on Medical Imaging*, vol. 19, no. 6, pp. 568–576, 2000.

[225] N. Passat, C. Ronse, J. Baruthio, J.-P. Armspach, and J. Foucher, "Watershed and multimodal data for brain vessel segmentation: Application to the superior sagittal sinus," *Image & Vision Computing*, vol. 25, no. 512-527, 2007.

[226] D. Chillet, J. Jomier, D. Cool, , and S. Aylward, "Vascular atlas formation using a vessel-to-image affine registration method," in *Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 2878 of *Lecture Notes in Computer Science*, pp. 335–342, 2003.

[227] D. Cool, D. Chillet, J. Kim, J.-P. Guyon, M. Foskey, and S. Aylward, "Tissue-based affine registration of brain images to form a vascular density atlas," in *Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 2879 of *Lecture Notes in Computer Science*, pp. 9–15, 2003.

[228] H. Bogunović, J. M. Pozo, R. Cárdenes, L. San Román, and A. F. Frangi, "Anatomical labeling of the circle of Willis using maximum a posteriori probability estimation," *IEEE Transactions on Medical Imaging*, vol. 32, no. 9, pp. 1587–1599, 2013.

[229] M. S. Ugurel, B. Battal, U. Bozlar, M. S. Nural, M. Tasar, F. Ors, M. Saglam, and I. Karademir, "Anatomical variations of hepatic arterial system, coeliac trunk and renal arteries: An analysis with multidetector CT angiography," *British Journal of Radiology*, vol. 83, no. 992, pp. 661–667, 2010.

[230] H. A. Alsaif, W. S. Ramadan, and S. Arabia, "An anatomical study of the aortic arch variations," *Journal of King Abdulaziz University – Medical Sciences*, vol. 17, no. 2, pp. 37–54, 2010.

[231] K. Ducksoo, D. E. Orron, and J. J. Skillman, "Surgical significance of popliteal arterial variants," *Annals of Surgery*, vol. 210, no. 6, pp. 776–781, 1989.

[232] S. Murphy, *Medical Image Segmentation in Volumetric CT and MR Images*. PhD thesis, University of Edinburgh, 2012.

[233] A. P. Avolio, "Multi-branched model of the human arterial system," *Medical and Biological Engineering and Computing*, vol. 18, no. 6, pp. 709–718, 1980.

[234] H. Kahraman, M. Ozaydin, E. Varol, S. M. Aslan, A. Dogan, A. Altinbas, M. Demir, O. Gedikli, G. Acar, and O. Ergene, "The diameters of the aorta and its major branches in patients with isolated coronary artery ectasia," *Texas Heart Institute Journal*, vol. 33, no. 4, pp. 463–468, 2006.

[235] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.

[236] A. Rosenfeld and P. de la Torre, "Histogram concavity analysis as an aid in threshold selection," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 13, no. 2, pp. 231–235, 1983.

[237] M. Schaap, *Quantitative Image Analysis in Cardiac CT Angiography*. PhD thesis, Erasmus University Rotterdam, 2010.

[238] T. Brox, B. Rosenhaum, D. Cremers, and H.-P. Seidel, "Nonparametric density estimation with adaptive, anisotropic kernels for human motion tracking," in *Proc. 2nd Human Motion Workshop*, vol. 4814 of *Lecture Notes in Computer Science*, pp. 152–165, 2007.

[239] D. Liu, K. S. Zhou, D. Bernhardt, and D. Comaniciu, "Search strategies for multiple landmark detection by submodular maximization," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2831–2838, 2010.

[240] M. Dikmen, Y. Zhan, and X. S. Zhou, "Joint detection and localization of multiple anatomical landmarks through learning," in *Medical Imaging: Computer-Aided Diagnosis*, vol. 6915 of *Proc. SPIE*, p. 691538, 2008.

[241] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.